bidt Digital Transformation Research Conference

19–20 November 2026
BAdW

Extended Programme: bidt Digital Transformation Research Conference 2026

bidt – Bavarian Research Institute for Digital Transformation
en.bidt.digital

bidt Digital Transformation Research Conference

Inhalt / Content

Übersicht / Overview – 19. November 2026 6

Übersicht / Overview – 20. November 2026 7

OPENING KEYNOTE
Hype Resistant? Research in Digital Times
Sabine Pfeiffer | bidt & FAU Erlangen-Nuremberg 8

KEYNOTE
The Meat Proxy – On the (Im)Possibility of Appropriate Reliance in AI Advice
Niklas Kühl | University of Bayreuth 8

KEYNOTE
The Right Amount of Wrong – Privacy, Accuracy, and Public Statistics
Simson Garfinkel | BasisTech, USA 8

WIRTSCHAFT & ARBEIT Session 1
Governance digitaler Transformation als betriebliche Gestaltungsaufgabe: zwischen Beteiligung, Aushandlung und organisationaler Handlungsfähigkeit 11

Die Logik schöner Fälle: Zur wechselseitigen Dynamik von Sinnstiftungsprozessen und Technik in organisationalen Digitalisierungsprojekten
René Werner | Johannes-Kepler-Universität Linz, Österreich 11

Wer gestaltet die Digitale Transformation? Zum Verhältnis direkter Partizipation und verfasster Mitbestimmung
Anna Kiemer | bidt 14

Betriebliche Aushandlung der Nutzung generativer „Alltags-KI“ – Eine Herausforderung für betriebliche Governance und Mitbestimmung
Julius Glaser | Georg-August-Universität Göttingen; Thomas Hardwig | Georg-August-Universität Göttingen 18

Governance digitaler Transformation zwischen Technologie und Handlungsfähigkeit: Eine Fallstudie zur Implementierung von Transformationskompetenz in einem Maschinenbauunternehmen
Kendra Pöhlmann | OTH Regensburg; Ruth Nürnberger | Brückner Maschinenbau 21

COMMUNICATION & SOCIETY Session 1
Digital Public Sphere: Conditions and Framework for a Constructive Debating Culture 26

AI-supported stakeholder simulation as a tool for digital participation: The Eckernkamp Compass
Nicholas H. Müller | SHIFT; Michael Müßig | SHIFT; Markus Oermann | SHIFT; Marie Herz | SHIFT; Hakan Arda | SHIFT; Julian Sehne | SHIFT 26

Platforms, Messengers and Messages: What Shapes Deliberative Engagement in Climate Scientists' Social Media Debates?
Stephan Schlögl | University of Passau; Paula Kolhep | University of Passau; Hannah Schmid-Petri | University of Passau & bidt 29

Hey ChatGPT, should I quit my job? – How the complexity of conversations with GenAI affects “wise” reasoning and enables human-AI synergy
Christina M. Brunner | LMU Munich; Franziska L. Salzberger | LMU Munich; Katharina G. Kugler | LMU Munich; Felix C. Brodbeck | LMU Munich 32

RECHT & ETHIK Session 1
Verantwortung und Kontrolle in KI-vermittelten Entscheidungen 37

Algorithmisches Ermessen: Verantwortung, Transparenz und Legitimität beim Einsatz von KI in der öffentlichen Verwaltung
Laura Crompton | TH Ingolstadt; Josefine Ruppenthal | TH Ingolstadt 37

Menschliche Aufsicht von AI Agents nach dem AI Act
Chiara Streitbörger | Universität Potsdam; Marietta Ostendorf | Universität Potsdam 42

Das Problem der Verantwortung in Zeiten KI-basierter Herrschaft
Juan Camilo Castillo Rivadeneira | InfAI; Vanita Römer | InfAI 49

WIRTSCHAFT & ARBEIT Session 2
Governance digitaler Transformation als Mehrebenen-Phänomen: zwischen Organisationen, Institutionen und Wertschöpfungssystemen 53

Transformation durch Interfaces: Digitale Wertschöpfungsökosysteme und die Verarbeitung pluraler gesellschaftlicher Anforderungen
Benjamin Doubali | Johannes Gutenberg Universität Mainz 53

Digitale Souveränität: Die nächste Herausforderung der digitalen Transformation
Walter Brenner | Universität St. Gallen, Schweiz; Barbara Brenner | Universität St. Gallen, Schweiz; Benjamin van Giffen | Universität Liechtenstein, Liechtenstein 57

Governance als Enabler: Menschliche Adaption als zentrale, vernachlässigte Steuerungsdimension digitaler Transformation
Imme Baumüller | Unabhängige Forscherin und Unternehmensberaterin, Digitale Transformation 58

COMMUNICATION & SOCIETY Session 2
Digital Public Sphere: Conditions and Framework for a Constructive Debating Culture 64

Relational Truths and Synthetic Publics: Political Communication in the Age of Generative AI
Ignas Kalpokas | Vytautas Magnus University, Lithuania 64

When Transparency Teaches: Explainable AI, Social Learning, and Evasion in Automated Content Moderation
Malika Talgatova | TU Munich; Yannis Theocharis | TU Munich; Nils Weidmann | University of Konstanz; Molly Roberts | University of California San Diego, USA 68

Humanizing Generative AI: Anthropomorphic Interpretations, Trust and Communicative Agency
Mohammad Hossein Jarrahi | University of North Carolina at Chapel Hill, USA; Amir Karami | Kennesaw State University, USA; Lisa Park | University of Alabama at Birmingham, USA; Ensieh Jarrahi | Independent Scholar; Christoph Lutz | BI Norwegian Business School, Norway 72

Strategic Knowledge as a Key Competency in Learning with Generative AI: Status Quo among Higher Education Students
Nick Naujoks-Schober | Ansbach University of Applied Sciences; Marion Händel | Ansbach University of Applied Sciences 75

LAW & ETHICS Session 2
Regulating Agentic AI: Autonomy, Risk, and Design 76

High Risk by Design? Agentic AI and Purpose Based Classification in the EU AI Act
Jülide Bredée | Capgemini Invent; Daniella Domokos | Capgemini Invent 76

What Should an AI Agent Be Permitted to Do on Its Own?
Damir Safin | fortiss; Dian Balta | fortiss; Timon Sengewald | FAU Erlangen-Nuremberg; Angela Roth | FAU Erlangen-Nuremberg; Hannah Lenk | University of Bamberg; Marc Redepenning | University of Bamberg 78

Risk, Principles, and Learning by Design: AI Regulatory Sandboxes and the Emerging EU Wide GPAI Sandbox under the AI Office
Fabio Seferi | IMT School for Advanced Studies Lucca & University of Florence, Italy; Armando Guio Espanol | Berkman Klein Center for Internet & Society, Harvard University, USA 82

The Right to an Independent AI Agent: Structural Separation in Agentic Commerce
Hofit Wasserman Rozen | Tel-Aviv University, Israel; Niva Elkin-Koren | Harvard University, USA 86

SUB-SESSION
Startups for State oder State versus Startups? Zwischen Innovationsdynamik und Verwaltungsrealität 87

SUB-SESSION
Understanding the Machine Heuristic: Trust, Credibility, and Bias in the Age of Generative AI 88

Effect of Machine Heuristic on Credibility of Generative AI as a Communication Source
S. Shyam Sundar | The Pennsylvania State University, USA 88

Generative AI: Biased but Trusted?
Markus Appel | University of Würzburg; Tanja Messingschlager | University of Würzburg 89

Trusting the Machine? How AI Authorship Shapes News Consumption
Daria Kravets-Meinke | bidt; Hannah Schmid-Petri | University of Passau & bidt 89

SUB-SESSION
A Digital Future Worth Living: From Ethical Principles to Practice in Software Engineering 90

Introduction
Klaus Staudacher | bidt & European Digital Humanism Initiative 90

Beyond AI Ethics: What Kind of Digital Future Do We Want? From Risk Mitigation to Human Flourishing
Erich Prem | eutema, TU Vienna & European Digital Humanism Initiative 91

From Principles to Practice: Can Digital Humanism Become a Design Paradigm? Embedding Human Values into Software Engineering
Alexander Pretschner | bidt & TU Munich 91

The Limits of AI Engineering: What Cannot Be Designed? Human Judgment, Responsibility, and Democratic Governance
Steven Umbrello | Institute for Ethics and Emerging Technologies, University of Turin & European Digital Humanism Initiative 91

ECONOMY & WORK Session 3
Governing Digital Transformation Under Tensions: Between Regulation, Organizational Dynamics, and New Dependencies 93

Regulating BigTech: YouTube's Response to Australia's Social Media Ban
Ina Partisch | Fraunhofer FIT / Frankfurt University of Applied Sciences; Michael Rosemann | Queensland University of Technology, Australia; Nadine Ostern | Queensland University of Technology, Australia 93

Are We Now Being Surveilled? Organizational Ignorance in Digitalized Waste Management
Alena Bleicher | Harz University of Applied Sciences; Sabine Biedermann Camposano | Harz University of Applied Sciences & TU Munich; Diana Ayeh | Helmholtz Centre for Environmental Research 97

Beyond Individual Overreliance: Generative AI Dependence, Workplace Norms and the Governance of Everyday Automation
Christoph Lutz | BI Norwegian Business School, Norway; Gemma Newlands | University of Zurich, Switzerland; Shubin Yu | HEC Paris, France 99

LAW & ETHICS Session 3
Responsibility and Control in AI-Mediated Decision-Making 103

Access to Digital Information as Infrastructure Relevant to Fundamental Rights – How Far Does State Responsibility Extend?
Charlotte Werther | Open Search Foundation e.V. 103

Democratic Legitimacy under the Inferential Turn: Epistemic Power, Traceability Gaps, and Institutional Reconstructability in AI-mediated Public Administration
Thomas Absmayr | FHV - Vorarlberg University of Applied Sciences, Austria 106

Statistical Lives, Identifiable Persons, and the Ethics of Predictive Policing
Korbinian Rüger | LMU Munich 111

SUB-SESSION
Gute wissenschaftliche Praxis als Governance-Problem: Wissenschaft und Lehre im Zeitalter generativer KI 114

SUB-SESSION
Zwischen Code und KI: Wandel der Arbeit in der Softwareentwicklung 115

Die Automatisierung der Automatisierung? Generative KI und die Refiguration von Programmierarbeit
Florian Butollo | Goethe-Universität Frankfurt; Ann Katzinski | WZB & Weizenbaum-Institut; Marlene Kulla | WZB/Weizenbaum-Institut 115

Frisst die KI ihre Erschaffer? Was KI mit der Arbeit in Software-Entwicklung und Programmierung macht
Sabine Pfeiffer | bidt & FAU Erlangen-Nürnberg; Marco Blank | FAU Erlangen-Nürnberg; Angela Graf | bidt 116

KI, Automatisierung und Arbeitsqualität in der Softwareentwicklung
Martin Krzywdzinski | WZB/Weizenbaum-Institut/Helmut-Schmidt-Universität Hamburg 116

SUB-SESSION
Generative KI als Interventionistin und Sparringspartnerin: Interaktionsdynamiken in digitalen Öffentlichkeiten verstehen und stabilisieren 117

Generative KI und die Neuordnung digitaler Öffentlichkeit
Andreas Jungherr | bidt & Universität Bamberg 118

KI als Trainerin: Konflikthafte Gespräche simulieren und reflektieren
Katharina V. Hajek | LMU München 118

KI als Interventionistin: KI-gestützte Interventionen in digitalen Öffentlichkeiten
Steliyana Doseva | bidt 118

SUB-SESSION
Interdisciplinarity for the Study of Socially Relevant Imagery 120

What Media Frames Reveal About Stance: A Dataset and Study of Memes in Climate Change Discourse
Shijia Zhou | Center for Information and Language Processing (CIS), LMU Munich 120

Conceptualizing and Detecting Toxic Entertainment in Visual Social Media Content
Ursula K. Schmid | LMU Munich; Philipp Spielbauer | LMU Munich 120

Where Have All the Memes Gone? Analyzing the Prevalence and Virality of Memes Using Vision-Language Models
Nadezhda Ozornina | Department of Media and Communication, LMU Munich 121

SUB-SESSION
Digital Competences: Insights and Strategic Implications for the Empowerment of Society for the Digital Future 122

The DigComp 3.0 update: Why, How, and the Framework's Potential for Defining and Assessing Digital Skills
Jude Cosgrove | Joint Research Centre, European Commission 123

Measuring and Status Quo of Digital Literacy Among Older Adults
Linda Göbl | Catholic University of Applied Sciences Freiburg; Michael Doh | Catholic University of Applied Sciences Freiburg 123

Digital Competences Around the Globe: A Comparative Study of Nine Countries
Roland A. Stürz | bidt; Danilo Harles | bidt 124

From Frameworks to Action: Operationalising and Measuring Digital Competence for Skills Policy and Workforce Development
Ulrike Domany-Funtan | Secretary General fit4internet, Austria 124

SUB-SESSION
Overcoming Key Challenges of Data Donations: Balancing Individual Rights and Collective Benefit 127

SUB-SESSION
Strafbare Deepfakes? Antworten auf neue Missbrauchsphänomene 131

SUB-SESSION
Führung und Organisationen im digitalen Wandel: Chancen, Herausforderungen und KI-Perspektiven 134

(Negatives) Führungsverhalten in digitalen Arbeitswelten
Ellen Schmid | Universität der Bundeswehr München 134

KI als Entscheidungshilfe für Führungskräfte?
Markus Böhm | Hochschule Landshut 134

Relationaler Wandel und soziale Dynamiken im KI-Zeitalter
Julia Reif | Universität der Bundeswehr München 135

SUB-SESSION
KI statt Streife und Richterbank? Gesellschaftliche Akzeptanz und Vertrauen in den Einsatz digitaler Technologien bei Polizei und Justiz 136

Referent:innen-Liste / List of Presenters 139

Conference Organisation 143

Conference Chairs 143

Organising Team 143

Programme Committees 143

bidt Digital Transformation Research Conference

Übersicht / Overview – 19. November 2026

08:30–09:00 h	Registrierung / Registration
09:00–09:15 h	Grußworte / Welcome · Conference Chairs
09:15–10:00 h	Eröffnungs-Keynote / Opening Keynote: Hype Resistant? Research in Digital Times · Sabine Pfeiffer | bidt & FAU Erlangen-Nuremberg
10:00–10:15 h	Kaffeepause / Coffee Break

10:15–12:15 h · Sessions 1 · parallel / In Parallel

WIRTSCHAFT & ARBEIT / ECONOMY & WORK SESSION 1	KOMMUNIKATION & GESELLSCHAFT / COMMUNICATION & SOCIETY SESSION 1	RECHT & ETHIK / LAW & ETHICS SESSION 1
Governance digitaler Transformation als betriebliche Gestaltungsaufgabe: zwischen Beteiligung, Aushandlung und organisationaler Handlungsfähigkeit Sprache / Language: Deutsch | Raum / Room: Plenarsaal	**Digital Public Sphere: Conditions and Framework for a Constructive Debating Culture** Sprache / Language: English | Raum / Room: Sitzungssaal 2	**Verantwortung und Kontrolle in KI-vermittelten Entscheidungen** Sprache / Language: Deutsch | Raum / Room: Bibliothek

12:15–13:15 h	Mittagspause / Lunch Break

13:15–15:15 h · Sessions 2 · parallel / In Parallel

WIRTSCHAFT & ARBEIT / ECONOMY & WORK SESSION 2	KOMMUNIKATION & GESELLSCHAFT / COMMUNICATION & SOCIETY SESSION 2	RECHT & ETHIK / LAW & ETHICS SESSION 2
Governance digitaler Transformation als Mehrebenen-Phänomen: zwischen Organisationen, Institutionen und Wertschöpfungssystemen Sprache / Language: Deutsch | Raum / Room: Plenarsaal	**Digital Public Sphere: Conditions and Framework for a Constructive Debating Culture** Sprache / Language: English | Raum / Room: Sitzungssaal 2	**Regulating Agentic AI: Autonomy, Risk, and Design** Sprache / Language: English | Raum / Room: Bibliothek

15:15–15:30 h	Kaffeepause / Coffee Break

15:30–17:00 h · Sub-Sessions · parallel / In Parallel

SUB-SESSION	SUB-SESSION	SUB-SESSION
Startups for State oder State versus Startups? Zwischen Innovationsdynamik und Verwaltungsrealität Sprache / Language: Deutsch | Raum / Room: Plenarsaal	**Understanding the Machine Heuristic: Trust, Credibility, and Bias in the Age of Generative AI** Sprache / Language: English | Raum / Room: Sitzungssaal 2	**A Digital Future Worth Living: From Ethical Principles to Practice in Software Engineering** Sprache / Language: English | Raum / Room: Bibliothek

17:00–18:00 h	Keynote: The Meat Proxy – On the (Im)Possibility of Appropriate Reliance in AI Advice · Niklas Kühl | University of Bayreuth
18:00–20:00 h	Get-Together

bidt Digital Transformation Research Conference

Übersicht / Overview – 20. November 2026

08:30–09:00 h	Registrierung / Registration

09:00–10:30 h · Sessions 3 & Sub-Session · parallel / In Parallel

WIRTSCHAFT & ARBEIT / ECONOMY & WORK SESSION 3 **Governing Digital Transformation Under Tensions: Between Regulation, Organizational Dynamics, and New Dependencies** Sprache / Language: English | Raum / Room: Plenarsaal	SUB-SESSION **Gute wissenschaftliche Praxis als Governance-Problem: Wissenschaft und Lehre im Zeitalter generativer KI** Sprache / Language: Deutsch | Raum / Room: Sitzungssaal 2	RECHT & ETHIK / LAW & ETHICS SESSION 3 **Responsibility and Control in AI-Mediated Decision-Making** Sprache / Language: English | Raum / Room: Bibliothek

10:30–11:30 h	Keynote: The Right Amount of Wrong – Privacy, Accuracy, and Public Statistics · Simson Garfinkel | BasisTech, USA
11:30–12:30 h	Mittagspause / Lunch Break

12:30–14:00 h · Sub-Sessions · parallel / In Parallel

SUB-SESSION **Zwischen Code und KI: Wandel der Arbeit in der Softwareentwicklung** Sprache / Language: Deutsch | Raum / Room: Plenarsaal	SUB-SESSION **Generative KI als Interventionistin und Sparringspartnerin: Interaktionsdynamiken in digitalen Öffentlichkeiten verstehen und stabilisieren** Sprache / Language: Deutsch | Raum / Room: Sitzungssaal 2	SUB-SESSION **Interdisciplinarity for the Study of Socially Relevant Imagery** Sprache / Language: English | Raum / Room: Bibliothek

14:00–15:30 h · Sub-Sessions · parallel / In Parallel

SUB-SESSION **Digital Competences: Insights and Strategic Implications for the Empowerment of Society for the Digital Future** Sprache / Language: English | Raum / Room: Plenarsaal	SUB-SESSION **Overcoming Key Challenges of Data Donations: Balancing Individual Rights and Collective Benefit** Sprache / Language: English | Raum / Room: Sitzungssaal 2	SUB-SESSION **Strafbare Deepfakes? Antworten auf neue Missbrauchsphänomene** Sprache / Language: Deutsch | Raum / Room: Bibliothek

15:30–15:45 h	Kaffeepause / Coffee Break

15:45–17:15 h · Sub-Sessions · parallel / In Parallel

SUB-SESSION **Führung und Organisationen im digitalen Wandel: Chancen, Herausforderungen und KI-Perspektiven** Sprache / Language: Deutsch | Raum / Room: Bibliothek	SUB-SESSION **KI statt Streife und Richterbank? Gesellschaftliche Akzeptanz und Vertrauen in den Einsatz digitaler Technologien bei Polizei und Justiz** Sprache / Language: Deutsch | Raum / Room: Sitzungssaal 2

17:15–17:30 h	Abschluss / Closing
ab / from 17:30 h	Ausklang / Networking

OPENING KEYNOTE
Hype Resistant? Research in Digital Times

Sabine Pfeiffer | bidt & FAU Erlangen-Nuremberg

19.11.2026, 09:15-10:00
Language: English

Digital transformation is accompanied by powerful narratives: of inevitable disruption, universal enthusiasm, or widespread fear. Research should not simply reproduce these narratives, but critically examine the assumptions, interests, and blind spots they contain. Yet this task is complicated by research funding itself. By privileging fashionable topics and promising rapid relevance, funding programmes can encourage researchers to follow the hype rather than question it. Maintaining intellectual independence therefore also means, at times, declining to chase the latest agenda.

Hype-resistant research also requires interdisciplinarity on equal terms. The social sciences cannot confine themselves to issuing warnings from the sidelines; they must engage more deeply with technologies in order to recognise both possibilities and limits of design. Conversely, computer science must not reduce the social sciences to securing user acceptance after the fact. It must take people seriously within its own concepts, methods, and development processes.

Drawing on my own empirical research, the keynote questions two common assumptions. First, does digitalisation actually feel as transformative and disruptive to people as public discourse suggests? Second, are people really predominantly afraid of digitalisation? Both examples illustrate why researchers should investigate rather than presume the experiences attributed to society—and why questioning the hype is not a rejection of innovation, but a precondition for understanding and shaping digital change responsibly.

KEYNOTE
The Meat Proxy – On the (Im)Possibility of Appropriate Reliance in AI Advice

Niklas Kühl | University of Bayreuth

19.11.2026, 17:00-18:00
Language: English

In 1995, Homer Simpson automated his safety job at the nuclear plant with a drinking bird. What was a joke then is now law: Article 14 of the EU AI Act requires effective human oversight of high-risk systems. But the question is not whether a human sits in the loop; it is whether that person can actually disagree, i.e., if they have the means to appropriately rely. Drawing on studies with 136 lay and expert decision-makers and with 101 radiologists, this keynote shows that explanations only help when their format invites verification, and that appropriate reliance and effective human oversight are a design property of the system, not a checkbox in a process.

KEYNOTE
The Right Amount of Wrong – Privacy, Accuracy, and Public Statistics

Simson Garfinkel | BasisTech, USA

20.11.2026, 10:30-11:30
Language: English

Differential privacy makes it possible to publish useful statistics about individuals in a group while limiting the quality of the inferences that can be made about any specific individual.

The U.S. Census Bureau first used DP in OnTheMap in 2008; the same protected employment data underlie OnTheMap for Emergency Management, launched in 2010.
DP makes it possible to control privacy loss, but DP does not recommend an acceptable balance between confidentiality and utility: that is left as a job for policymakers.
DP's application to the 2020 Census generated controversy and resulted in numerous lawsuits, attacking the accuracy and legality of its use. These technical disputes became entangled with political claims that deliberately protected statistics were falsified.
In June 2026, the US Commerce Department issued Department Administrative Order (DAO) 216-26, "Disclosure Avoidance for Statistical Products," which banned the use of differential privacy and other approaches for adding noise to protect privacy for any statistical produce disseminated by the bureaus of the Office of the Under Secretary for Economic Affairs, including the US Census Bureau and the US Bureau for Economic Analysis.
The prohibition does not eliminate the legal requirement to protect confidential data. Instead, it directs agencies toward aggregation, coarsening, and suppression, potentially producing less detailed public information. The dispute illustrates a recurring conflict among political objectives, legal requirements, and scientific reality—most obviously here with respect to privacy protection, but which is present in other areas as well.

bidt Digital Transformation Research Conference

Sessions
Day 1

Thursday, 19.11.2026

WIRTSCHAFT & ARBEIT Session 1
Governance digitaler Transformation als betriebliche Gestaltungsaufgabe: zwischen Beteiligung, Aushandlung und organisationaler Handlungsfähigkeit

19.11.2026, 10:15-12:15

Sprache: Deutsch

Programmkomitee: Angela Graf | bidt; Sabine Pfeiffer | bidt & FAU Erlangen-Nürnberg; Thomas Hess | bidt & LMU München; Alexander Pretschner | bidt & TU München

Moderation: Angela Graf | bidt

Die Logik schöner Fälle: Zur wechselseitigen Dynamik von Sinnstiftungsprozessen und Technik in organisationalen Digitalisierungsprojekten

René Werner | Johannes-Kepler-Universität Linz, Österreich

Die Versprechungen für Künstliche Intelligenz (KI) sind in der medialen Darstellung nahezu grenzenlos. Vom Traum vollständiger Automatisierung ‚unliebsamer' Tätigkeiten zur industriellen Revolution des 21. Jahrhunderts oder sogar dem Erreichen von allgemeiner Künstlicher Intelligenz, die gesellschaftlichen Narrative über KI sind allumfassend (Hirsch-Kreinsen & Krokowski, 2023: 477-478; Sudman 2018: 59-60). Solchen Erzählungen einer digitalen Transformation der ganzen Gesellschaft setze ich mit diesem Beitrag einen nüchternen Blick entgegen, in dem ich auf Basis meiner noch unveröffentlichten Dissertation in der Soziologie der Fragestellung nachgehe, wie Organisationsmitglieder organisationale Praxis mit KI vollziehen. Diese Frage zielt also nicht auf eine diskursive Ebene, sondern betrachtet einen bestimmten sozialen Kontext der Einführung und Nutzung von KI als spezifischer digitaler Technologie: Die Transformation von Arbeit in Organisationen. Ein zentrales Ergebnis meiner Dissertation zu diesem Gegenstand besteht darin, dass die Organisationsmitglieder, deren Arbeit mit KI teilweise automatisiert werden soll, sehr wohl auch ein aktives Gestaltungspotenzial nutzen können und dies auch praktisch tun. Dabei zeitigen sich jedoch auch nicht-intendierte Folgen, die sich sowohl auf die Funktionsweise des eingeführten KI-Systems, aber vor allem auch auf die eigene Arbeitspraxis dieser Akteur:innen auswirken. Mein Beitrag argumentiert in diesem Sinn dafür, dass eine Governance von digitalen Transformationsprozessen zwangsläufig reflexiv und anpassungsfähig erfolgen muss, selbst wenn die Betroffenen der Technologieeinführung keine passiven Empfänger:innen von hierarchisch getroffenen Entscheidungen in der Organisation sind. Dafür kombiniere ich konzeptionell Sensemaking nach Karl Weick (1995; Weick et al. 2005) mit einer praxistheoretischen Fassung der Nutzung von Technik in Organisationen (Nicolini 2012) und erarbeite diese Konzeptualisierung anhand einer qualitativen Einzelfallstudie zur Einführung von KI in einer Organisation.

Empirisch betrachte ich einen Fall der Einführung eines KI-Systems in eine deutschsprachige Krankenversicherung. Das Ziel der Einführung ist es den fallbasierten organisationalen Prozess der Kostenerstattung, der von Sachbearbeiter:innen durchgeführt wird, zunächst teilweise, aber in die Zukunft gerichtet auch größtenteils zu automatisieren. Sachbearbeiter:innen entscheiden für jede erbrachte medizinische Leistung, inwiefern diese ein Versicherungsfall darstellt, der von der Krankenversicherung übernommen wird oder aber ganz oder teilweise von den Versicherten selbst erbracht werden muss. Dafür verwenden die Sachbearbeiter:innen Expert:innenwissen über Tarifkataloge, die Krankenversicherungen mit ärztlichen Praxen geschlossen haben. Das KI-System soll diesen Prozess automatisieren, in dem die eingereichten Dokumente auf die wichtigen Informationen wie, Arztpraxis, Versicherte, medizinische Leistung, Kosten etc. ausgelesen und dann in das zentrale Verwaltungssystem überführt werden. Diesen Prozess habe ich über mehrere Jahre in einer Einzelfallstudie (Siggelkow 2007) untersucht, in der ich auf verschiedene qualitative Datentypen zugegriffen habe: Expert:inneninterviews mit unterschiedlichen Akteursgruppen in

diesem Prozess wie Sachbearbeiter:innen, Gruppenleitungen, Projektverantwortlichen und dem KI-Entwicklungsteam an verschiedenen Standorten der Krankenversicherung. Diese Daten habe ich ergänzt um Dokumentenanalysen zum KI-System und dem Prozess der Kostenerstattung sowie um teilnehmende Beobachtungen an den Standorten und den wöchentlichen Projektmeetings. Basierend auf diesen Datentypen habe ich mit der qualitativen Inhaltsanalyse nach Gläser und Laudel (2010) ein Modell der Dynamik der Wechselseitigkeit von organisationaler Praxis und digitaler Technologie entwickelt.

Theoretisch kombiniere ich zwei Prozessperspektiven miteinander: Sensemaking nach Weick ist eine Prozessperspektive auf Organisationen, welche die Sinndeutungen der Akteure in den Vordergrund stellt (Sandberg & Tsoukas 2015: 9ff.). Diese Sinnstiftungen sind nicht beliebig, sondern stark durch den sozialen Kontext geprägt, zum Beispiel dem beruflichen Kontext der Arbeit in der Kostenerstattung einer Krankenversicherung. Historisch gesehen wurde Weicks Sensemaking zunächst sehr kognitivistisch gedacht und danach - auch von Weick selbst - stärker sozialkonstruktivistisch (Weick et al. 2005: 409). Sinn wird also nicht nur individuell mental erzeugt; in sozialen Kontexten sind die Sinnstiftungen das Ergebnis einer ständig laufenden Aushandlung mit anderen Akteur:innen. Eine solche Lesart von Sensemaking kombiniere ich mit einer praxistheoretischen Betrachtung von der Nutzung von Technik in Organisationen. Basierend auf Nicolinis Arbeiten (2012) betone ich damit auch die Materialität von Technik, wie etwa dem eingeführten KI-System und den iterativen Character der Nutzung eines solchen im Kontext von Kostenerstattung als organisationaler Praxis (vgl. Nicolini 2012; Leonardi 2011: 3ff.). Die Kombination von Sensemaking und Praxistheorie ermöglicht es mit daher die Sinnstiftungen der organisationalen Akteure ernst zu nehmen, die maßgeblich den Umgang mit neu eingeführter Technik prägen, aber auch gleichzeitig nicht nur auf die Akteure fokussiert zu sein, sondern auch die materiellen Besonderheiten der eingeführten Technik. Meine Perspektive ist dabei ebenso sensibel für die sozialen Aushandlungsprozesse, die mit der Einführung von neuer Technik in Organisationen einhergehen, auch was die Rolle von Macht und Konflikten angeht. Diese Aushandlung und in diesem Sinne Governance der Technologieeinführung verläuft nicht nur anhand von organisationalen Hierarchien. Die Sachbearbeiter:innen können ihre Expertise in Form von kommunizierten Sinnstiftungen durchaus auch als Ressource im Machtkampf dieses Digitalisierungsvorhabens einführen (Schildt et al. 259ff.).

Meine Analyse dieses Falls im Verhältnis dieser Fragestellung, theoretischen Perspektive und methodischen Designs betont zwei zentrale Ergebnisse: Erstens sind die Sinnstiftungen zu ihrer Arbeit und Nutzung des KI-Systems durch die Sachbearbeiter:innen als organisationale Akteure zentral um die Aushandlungen um diesen Prozess der angestrebten Automatisierung von Arbeit zu verstehen. Sachbearbeiter:innen weisen eine durchaus ästhetische Dimension (vgl. Strati 1996; 2010) ihrer Arbeit aus: Die Logik schöner Fälle. Schöne Fälle sind diejenigen, die Sachbearbeiter:innen auf einen ersten Blick bewerten können. Das bedeutet nicht unbedingt einfache Fälle, jedoch solche, in denen alle notwendigen Informationen auf einen Blick verfügbar sind. Dieser Blick wird dabei auch in der Ausbildung neuer Sachbearbeiter:innen erlernt und stellt eine verkörperte Praxis dar. Diese Logik schöner Fälle wird auch als zentrale Ressource genutzt um den Prozess der Einführung des KI-Systems mit zu gestalten. In den Situationen, in denen das KI-System nicht mit der Einschätzung der Sachbearbeiter:innen zusammenfällt, wird eine Sinnstiftung erzeugt, die das KI-System als defizitär betrachtet. Diese Sinnstiftung wird gegenüber anderen Akteursgruppen angeführt und auch gegen andere Sinnstiftungen verteidigt. Dies führt dazu, dass es zu Anpassungen im KI-System kommt. Das implizite Wissen, dass die Sachbearbeiter:innen über die Jahre aufgebaut haben, nämlich wie ärztliche Praxen entgegen des offiziellen Tarifkatalogs bestimmte Leistungen bezeichnen und die Rechnungen aufbauen, ist eine zentrale Fehlerquelle des KI-Systems. Mithilfe einer Liste an synonymen Bezeichnungen, welche die Sachbearbeiter:innen selbst erstellen und pflegen, soll das KI-System darauf trainiert werden, auch solche abweichenden Fälle automatisiert richtig zu erkennen.

Zweitens zeigt sich, dass sich im Zuge der Einführung von Technik auch nicht-intendierte Folgen zeigen können. Die Sachbearbeiter:innen nutzen ihre Expertise für ihre Arbeit als Ressource um die weitere

Gestaltung des KI-Systems mitzubestimmen. Dies hat jedoch den Effekt, dass der zentrale, verkörperte Wissensbestand der Logik schöner Fälle, im Verlauf des Prozesses zunehmend an Bedeutung verliert, da das KI-System genau diese Typen von Fällen relativ zuverlässig automatisiert bearbeiten kann. Dadurch kommt es zu einer Verdichtung und Verschiebung von Arbeit (vgl. Hardering et al. 2016: 8f.), da die Sachbearbeiter:innen nun im Alltag mehr mit komplexen und Sonderfällen zu tun haben, die auch eine andere Art von (Erfahrungs-)Wissen voraussetzen. Daraus folgen auch verschiedene Blicke auf die Frage, welchen Stellenwert Sachbearbeiter:innen zukünftig haben sollen: Werden Stellen abgebaut, weil nun viele Fälle automatisiert bearbeitet werden können oder erhalten Sachbearbeiter:innen nun sogar eine höhere Bezahlung, da sie zunehmend zu Expert:innen für Sonderfälle der Kostenerstattung werden?

Mein Beitrag betrachtet die Einführung von KI als einer spezifischen digitalen Technologie in einen bestimmten sozialen Kontext, nämlich Arbeit in Organisationen mithilfe einer praxistheoretischen Sensemaking-Perspektive. Diese Perspektive ist geeignet, die Besonderheiten von KI und vor allem auch die dynamischen und zeitsensitiven Aushandlungsprozesse zu analysieren, die eine solche Einführung von Technik in Organisationen prägen. Damit trage ich zum Wissensstand an der Schnittstelle von Arbeits-, Organisations- und Techniksoziologie und deren Betrachtung von digitaler Transformation bei.

Literatur

Gläser, J. & Laudel, G. (2010). Experteninterviews und qualitative Inhaltsanalyse: als Instrumente rekonstruierender Untersuchungen (4. Auflage). VS Verlag für Sozialwissenschaften.

Hardering, F., Will-Zocholl, M. & Hofmeister, H. (2016). Sinn der Arbeit und sinnvolle Arbeit: Zur Einführung. Arbeit, 24(12), 3–12. https://doi.org/10.1515/arbeit-2016-0002

Hirsch-Kreinsen, H. & Krokowski, T. (2023). Technologieversprechen Künstliche Intelligenz. Vergangene und gegenwärtige Konjunkturen in der Bundesrepublik. Berliner Journal für Soziologie, 33(4), 453–484. https://doi.org/10.1007/s11609-023-00504-1

Leonardi, P. M. (2011). When Flexible Routines Meet Flexible Technologies: Affordance, Constraint, and the Imbrication of Human and Material Agencies.

Nicolini, D. (2012). Practice theory, work, and organization: An introduction (1. Aufl.). Oxford University Press.

Sandberg, J. & Tsoukas, H. (2015). Making sense of the sensemaking perspective: Its constituents, limitations, and opportunities for further development. Journal of Organizational Behavior, 36(S1), 6-32. https://doi.org/10.1002/job.1937

Schildt, H., Mantere, S. & Cornelissen, J. (2020). Power in Sensemaking Processes. Organization Studies, 41(2), 241–265. https://doi.org/10.1177/0170840619847718

Siggelkow, N. (2007). Persuasion With Case Studies. Academy of Management Journal, 50(1), 20–24. https://doi.org/10.5465/amj.2007.24160882

Strati, A. (1996). Organizations Viewed through the Lens of Aesthetics. Organization, 3(2), 209–218. https://doi.org/10.1177/135050849632004

Strati, A. (2010). Aesthetic Understanding of Work and Organizational Life: Approaches and Research Developments. Sociology Compass, 4(10), 880–893. https://doi.org/10.1111/j.1751-9020.2010.00323.x

Sudmann, A. (2018b). Szenarien des Postdigitalen: Deep Learning als Medienrevolution. In A. Sudmann & C. Engemann (Hrsg.), Machine Learning - Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz (S. 55–73).

Weick, K. E. (1995). Sensemaking in Organizations (3. Auflage). Foundations for organizational science. Thousand Oask; SAGE.

Weick, K. E., Sutcliffe, K. M. & Obstfeld, D. (2005). Organizing and the Process of Sensemaking. Organization Science, 16(4), 409–421. https://doi.org/10.1287/orsc.1050.0133

Wer gestaltet die Digitale Transformation? Zum Verhältnis direkter Partizipation und verfasster Mitbestimmung

Anna Kiemer | bidt

Relevanz und Fragestellung

„Die digitale Transformation verändert die Arbeitswelt tiefgreifend. Im Zentrum dieser Veränderung stehen die Angestellten. Sie sind gleichermaßen Betroffene der Digitalisierung und diejenigen, die die Transformation maßgeblich gestalten“ (Kämpf und Lühr 2024, S. 34).

An diese Beobachtung knüpft der geplante Beitrag an. Denn wenn Beschäftigte digitale Transformation (DT) nicht nur erleben, sondern als Träger von Erfahrungswissen und technologischer Expertise praktisch mitgestalten, dann ist Beteiligung keine nachgelagerte Frage, sondern wesentliche Bedingung, dass die Transformation zustande kommt (z.B. Kane 2019). Zugleich ist sie eine zentrale arbeits- und organisationspolitisch relevante Frage von Handlungsmacht, Selbstwirksamkeit und Anerkennung (Dörre 2001; Gerbracht et al. 2024). Gerade aufgrund dieser doppelten Bedeutung von Beteiligung wird deutlich, dass DT mehr als eine technologische Veränderung ist. Vielmehr wird sie in der aktuellen Forschung als tiefgreifender organisationaler Wandel verstanden (z.B. Hess et al. 2016), der grundlegende Veränderungen auf struktureller, kultureller sowie auf Prozess- und Produktebene umfasst (Vial 2021). Dabei kommt es oft zu abrupten, widersprüchlichen Umbrüchen (Pfeiffer et al. 2024), die bestehende Logiken, Handlungsmuster und Routinen infrage stellen und Unsicherheit und Widerstand hervorrufen können (Graf et al. 2022; Hinings et al. 2018).

Gerade weil DT keinen eindeutigen Zielzustand besitzt, sondern als fortlaufend neu zu justierender Veränderungsprozess verläuft, stellt sich die Frage, wie solche Prozesse koordiniert und verantwortet werden können (Sciuk et al. 2025; Carroll et al. 2023). Damit rückt für den geplanten Beitrag die Frage in den Blick, wie sich die Beteiligung von Beschäftigten gestaltet und gestaltet werden kann. Beteiligung lässt sich insofern als eine spezifische Ordnungsdimension der Governance der DT begreifen: jene Ebene, auf der geregelt wird, wer an Transformationsprozessen beteiligt ist und in welcher Form Einfluss ausgeübt werden kann.

Dabei lässt sich zwischen direkter Partizipation im Arbeitsprozess, insbesondere im Kontext agiler und projektförmiger Arbeitsweisen (Haipeter et al. 2019; Kämpf und Lühr 2024) und institutionalisierter Mitbestimmung durch Betriebsräte und kollektive Interessenvertretungen unterscheiden (Müller-Jentsch 2022). Gerade diese Unterscheidung gewinn im Kontext der DT an Bedeutung, weil mit ihr neue Arbeits- und Organisationsformen einhergehen, in denen sich die Modi der Beteiligung verschieben. Im Zuge von Agilität, New Work oder holakratischen Ansätzen wird Partizipation stärker in den Arbeitsprozess verlagert, ohne dass damit automatisch erweiterte Einflussmöglichkeiten einhergehen (u.a. Nicklich et al. 2021). Dies erzeugt paradoxe Effekte: Beteiligung bleibt selektiv oder wird funktional gerahmt und schlägt in Mehrarbeit und Überforderung um (z.B. Carls, K., Gehrken, H., Kuhlmann, M., Splett, B., & Thamm, L 2023). Gleichzeitig werden diese neuen Partizipationsformen teils als Alternative verfasster Mitbestimmung konzipiert (Sattelberger et al. 2015), die im Zuge der DT selbst unter Druck gerät. So gelingt es betrieblichen Interessenvertretungen häufig nur eingeschränkt an die neuen Arbeitskontexte anzuschließen, da unklare Zuständigkeiten, begrenzte

Ressourcen und teils schwache Beteiligungsrechte ihre Einbindung erschweren (z.B. Rego 2022). Ob sich beide Beteiligungsformen schließlich ergänzen oder in Spannung zueinander stehen, ist dabei weiterhin umstritten (Boes et al. 2006).

Der vorliegende Beitrag setzt an dieser Stelle an. Er untersucht das Verhältnis von direkter Partizipation und verfasster Mitbestimmung als eine spezifische Ordnungsdimension der Governance von DT: Beide Formen werden dabei nicht als konkurrierende Alternativen verstanden, sondern als unterschiedliche, potenziell verschränkte Beteiligungsformen, deren Zusammenspiel empirisch rekonstruiert werden muss. Daraus ergibt sich folgende Fragestellung:

Wie werden direkte Partizipation und verfasste Mitbestimmung als Elemente der Governance digitaler Transformation praktisch hervorgebracht, institutionell gerahmt und zueinander in Beziehung gesetzt? Diese Frage soll im geplanten Beitrag am Beispiel der wissenschaftlichen Begleitung und Analyse eines interdisziplinären und crossfunktionalen Innovationsprojekts beantwortet werden.

Theoretischer/konzeptioneller Rahmen
Um das Auftreten und Verhältnis direkter Partizipation und verfasster Mitbestimmung im Kontext der DT zu analysieren, verknüpft der Beitrag strukturations- und praxistheoretische Perspektiven, die es erlauben, Handlung und Struktur als wechselseitig aufeinander bezogene Ebenen sozialer Praxis zu begreifen. In Anlehnung an Anthony Giddens wird soziales Handeln als rekursive Hervorbringung von Regeln und Ressourcen verstanden, die zugleich Bedingung und Ergebnis sozialer Praxis sind (Giddens 1988).

Aus dieser Perspektive lassen sich direkte und indirekte Partizipation als zwei verschränkte Ebenen sozialer Praxis fassen, die im Kontext der DT besonders relevant werden: direkte Partizipation als situative Hervorbringung und Organisation von Beteiligung im Arbeitsalltag in neuen projektförmigen, agilen Settings und indirekte Partizipation als institutionelle Rahmung der neuen Arbeitsformen durch formal abgesicherte Instrumente der Mitbestimmung. Beide stehen in einem rekursiven Verhältnis, insofern, dass direkte Partizipation auf institutionelle Voraussetzungen angewiesen ist, während indirekte Partizipation erst in konkreten Praxiszusammenhängen der DT wirksam wird (Kämpf und Lühr 2024).

Ergänzt wird dieser strukturationstheoretische Zugriff durch Orlikowskis practice lens, wonach Strukturen nicht einfach in Technologien oder Organisationen eingeschrieben sind, sondern in der wiederkehrenden Nutzung von Technologie praktisch hervorgebracht werden. Für den Beitrag ist das wichtig, weil damit die Dynamiken digitaler Transformationsprozesse insbesondere im konkreten Arbeitsalltag sichtbar werden, in denen Strukturen von Zuständigkeit, Entscheidung und Zusammenarbeit durch den Einsatz digitaler Technologien zugleich ermöglicht und begrenzt werden. Governance meint aus dieser Perspektive die Art und Weise, wie digitale Transformationsprozesse hervorgebracht und koordiniert werden. Partizipation und Mitbestimmung sind dabei zentrale Elemente über die die DT praktisch wirksam wird.

Methodische Vorgehensweise
Empirisch basiert der Beitrag auf einer qualitativen Fallstudie eines Maschinenbauunternehmens. Analysiert werden zwei empirisch miteinander verknüpfte Datensätze, die im selben Fallunternehmen unterschiedliche Beteiligungsordnungen im Kontext der DT erfassen: Interviews mit Betriebsratsmitgliedern sowie die Begleitung eines interdisziplinären, crossfunktionalen Innovationsprojekts.

Der empirische Zugriff greift damit die theoretische Perspektive auf Struktur und Handlung auf und untersucht einerseits die praktische Hervorbringung direkter Partizipation im Projekt und andererseits deren institutionelle Rahmung durch verfasste Mitbestimmung.
Das Projektmaterial, bestehend aus Interviews, Beobachtungen und internen Dokumenten, ermöglicht eine Rekonstruktion der Teamdynamiken, der Rollenbildung, der digitalen Koordination und der praktischen Hervorbringung projektförmiger, direkter Partizipation im Innovationsprojekt. Die Betriebsratsinterviews ergänzen diese Perspektive um die Sicht der verfassten Mitbestimmung auf die neuen Arbeits- und Beteiligungsformen im Innovationsprojekt und geben Einblicke in das Selbstverständnis, die Rolle und die Zuständigkeiten des Betriebsrats im Kontext der DT.

Zentrale Ergebnisse und Thesen
Im Innovationsprojekt wird sichtbar, dass sich im Zuge der DT neue Formen direkter Beteiligung herausbilden, die für den Projektverlauf zentral sind, zugleich aber selektiv bleiben und nur begrenzt institutionell gerahmt werden, während der Betriebsrat diese neuen Arbeits- und Beteiligungsformen nur am Rand als eigenen

Gestaltungsgegenstand aufgreift.

These 1: Im Innovationsprojekt entstehen neue direkte und informelle Formen der Partizipation
In Meetings, Abstimmungen, durch digitale Tools und in der konkreten Entwicklungsarbeit entstehen (neue) Rollen, Routinen, Kommunikationswege und Entscheidungslogiken. Die Zusammenarbeit ist dabei vergleichsweise offen, fachübergreifend und stark am Prinzip technischer Expertise orientiert. Entscheidungen werden häufig konsensorientiert getroffen und entlang von Zuständigkeiten ausgehandelt. Die Organisation dieser Beteiligung erfolgt jedoch maßgeblich über das Management und die Projektleitung, die Aufgaben und Prioritäten festlegen und damit den Rahmen setzen, innerhalb dessen direkte Partizipation möglich wird. Für die Governance der DT zeigt sich damit, dass Transformation wesentlich in neuen, arbeitsprozessnahen Beteiligungsformen wirksam wird, die zwar auf bestehenden organisationalen Strukturen aufbauen, aber zugleich eigenständige Modi der Koordination und Mitgestaltung hervorbringen.

These 2: Diese projektförmige Partizipation eröffnet Gestaltungsspielräume, bleibt jedoch selektiv und fragil
Trotz der offenen und kooperativen Zusammenarbeit sind Beteiligungsmöglichkeiten im Projektteam ungleich verteilt. Wer tatsächlich mitwirken kann, hängt stark von Ressourcen, zeitlicher Verfügbarkeit, Projektbindung und organisationalen Schnittstellen ab. Parallelprojekte, Reisen, funktionale Differenzierungen (z. B. zwischen Mechanik, Cloud und Vertrieb) sowie die Zuteilung von Aufgaben begrenzen die tatsächliche Beteiligung. Partizipation ist damit zwar „real“, aber weder für alle gleichermaßen zugänglich noch institutionell abgesichert. Für die Governance der DT folgt daraus, dass direkte Partizipation zwar eine zentrale Voraussetzung für die praktische Organisation und Umsetzung von Transformation darstellt, zugleich aber stark situativ und voraussetzungsvoll bleibt und sich nur begrenzt über das Projekt hinaus verstetigen kann.

These 3: Verfasste Mitbestimmung erfüllt im Kontext der DT weitgehend eine Schutz- statt Gestaltungsfunktion
Während im Projekt neue Formen direkter Partizipation entstehen, bleibt die institutionalisierte Mitbestimmung weitgehend auf Distanz. Der Betriebsrat betrachtet die projektförmige, digitale Produktentwicklung nicht als seinen Zuständigkeitsbereich, sondern bezieht sich vor allem auf formale Regeln, Verfahren und Schutzrechte und widmet sich eher anderen Themenfeldern wie beispielsweise Homeoffice-Regelungen. Damit entzieht sich der Betriebsrat weitgehend der operativen Gestaltung der DT, die vor allem durch das Management getrieben und durch direkte Partizipation der Beschäftigten im Arbeitsprozess getragen wird. Die zentrale Folge ist daher nicht das vollständige Fehlen von Mitbestimmung, sondern ihre funktionale Begrenzung: Sie bearbeitet eher die Folgen und Randbedingungen der DT als deren konkrete Ausgestaltung. Damit erfüllt institutionalisierte Mitbestimmung im Kontext der DT weitgehend eine Schutz- statt Gestaltungsfunktion.

Beitrag zur Governance digitaler Transformation
Der Beitrag trägt zur interdisziplinären Diskussion der Governance der DT bei, indem er die bislang oft getrennt betrachteten Perspektiven direkter Partizipation und verfasster Mitbestimmung empirisch zusammenführt und zeigt, dass sich die Qualität von Beteiligung nicht allein daran entscheidet, ob und in welchem Umfang Beschäftigte einbezogen werden, sondern daran, wie projektförmige Mitgestaltung und institutionalisierte Mitbestimmung aufeinander bezogen sind oder eben nicht. Der empirische Fall verweist dabei weniger auf einen Konflikt zwischen beiden Beteiligungsformen als auf eine strukturelle Entkopplung, in deren Folge die Gestaltung von Partizipation weitgehend in projektförmige, managementgetragene Strukturen verlagert wird, während institutionalisierte Mitbestimmung auf eine selektive Schutz- und Regelungsfunktion begrenzt bleibt. DT erscheint damit nicht als Ausweitung, sondern als Reorganisation von Beteiligung, in der Partizipation funktionalisiert und ungleich verteilt wird. Zugleich werden darin bislang ungenutzte Ansatzpunkte für betriebliche Interessenvertretung sichtbar, beispielsweise im Hinblick auf Überlastung durch

Beteiligungsanforderungen, selektive Teilhabechancen, erschwerte Kommunikation in digitalen Arbeitskontexten sowie Fragen von Anerkennung und Selbstwirksamkeit.

Literaturverzeichnis

Boes, Andreas; Kämpf, Tobias; Marrs, Kira (2006): Zukunft der Arbeitsbeziehungen: Zwischen Atomisierung und neuer Solidarität. In: *WSI Mitteilungen* 59 (2), S. 98–103.

Carls, K., Gehrken, H., Kuhlmann, M., Splett, B., & Thamm, L (2023): Digitalisierung aus Beschäftigtensicht. Fehlende Mitgestaltung, belastender Zusatzaufwand, mangelnde Wertschätzung. In: *WSI-Mitteilungen* 76 (2), S. 83–92.

Carroll, Noel; Hassan, Nik Rushdi; Junglas, Iris; Hess, Thomas; Morgan, Lorraine (2023): Transform or be transformed: the importance of research on managing and sustaining digital transformations: Taylor & Francis (32). In: *European Journal of Information Systems* (3).

Dörre, Klaus (2001): Partizipation im Arbeitsprozess—Alternative oder Ergänzung zur Mitbestimmung? In: *Industrielle Beziehungen. Zeitschrift für Arbeit, Organisation und Management*, S. 379–407.

Gerbracht, Marc; Krüger, Max; Castro Leal, Débora de; Tolmie, Peter; Wulf, Volker (2024): "We are no Luddites!" - CSCW, Co-Determination and Digital Transformation in Germany. In: *Proc. ACM Hum.-Comput. Interact.* 8 (CSCW1), S. 1–23. DOI: 10.1145/3637430.

Giddens, Anthony (1988): Die Konstitution der Gesellschaft: Campus Frankfurt/New York.

Graf, Angela; Hess, Thomas; Müller, Lea (2022): Digitale Transformation als Identitätsfrage. Zum Zusammenhang zwischen organisationaler Identität und digitaler Transformation am Beispiel eines Maschinenbauunternehmens. In: *AIS Studien* 15 (2), S. 44–61.

Hess, Thomas; Matt, Christian; Benlian, Alexander; Wiesböck, Florian (2016): Options for formulating a digital transformation strategy. In: *MIS Quarterly Executive* 15 (2), S. 103–119.

Kämpf, Tobias; Lühr, Thomas (2024): Angestellte und Mitbestimmung in der digitalen Transformation. In: *WSI-Mitteilungen* 77 (1), S. 34–41. DOI: 10.5771/0342-300X-2024-1-34.

Kane, Gerald (2019): The technology fallacy: people are the real key to digital transformation. In: *Research-Technology Management* 62 (6), S. 44–49.

Nicklich, Manuel; Sauer, Stefan; Pfeiffer, Sabine (2021): Antecedents and Consequences of Agility--On the Ongoing Invocation of Self-Organization. In: Sabine Pfeiffer, Manuel Nicklich und Stefan Sauer (Hg.): The Agile Imperative : Teams, Organizations and Society under Reconstruction? Cham: Springer International Publishing, S. 19–38.

Pfeiffer, Sabine; Nicklich, Manuel; Schreyer, Jasmin (2024): Digitalisierung der Arbeitswelten–die systemische Transformation verstehen. Ein Ausblick. In: Sabine Pfeiffer, Manuel Nicklich, Michael Henke, Martina Heßler, Martin Krzywdzinski und Ingo Schulz-Schaeffner (Hg.): Digitalisierung der Arbeitswelten: Zur Erfassbarkeit einer systemischen Transformation: Springer, S. 559–584.

Rego, Kerstin (2022): Works councils and the digitalisation of manufacturing: Opportunity or threat for their power position? In: *Economic and Industrial Democracy* 43 (4), S. 1911–1933.

Sattelberger, Thomas; Welpe, Isabell M.; Boes, Andreas (2015): Das demokratische Unternehmen: neue Arbeits- und Führungskulturen im Zeitalter digitaler Wirtschaft. 1. Auflage. Freiburg, München: Haufe Gruppe. Online verfügbar unter https://shop.haufe.de/prod/das-demokratische-unternehmen.

Sciuk, Christian; Engert, Simon; Gierlich-Joas, Maren; Hess, Thomas (2025): How Companies Navigate the (Un) Charted Waters of Digital Transformation. In: *California Management Review* 68 (1), S. 5–31.

Vial, Gregory (2021): Understanding digital transformation. A review and a research agenda. In: Andreas Hinterhuber, Tiziano Vescovi und Francesca Checchinato (Hg.): Managing digital transformation. Understanding the strategic process. London: Routledge, S. 13–66.

Betriebliche Aushandlung der Nutzung generativer „Alltags-KI“ – Eine Herausforderung für betriebliche Governance und Mitbestimmung

Julius Glaser | Georg-August-Universität Göttingen; Thomas Hardwig | Georg-August-Universität Göttingen

Ohne Wissen des Arbeitgebers nutzten 2024 29% der Befragten privat verfügbare Anwendungen mit generativer künstlicher Intelligenz (genKI) wie ChatGPT, Mistral, Midjourney usw. am Arbeitsplatz (Pfeiffer, 2025). „Alltags-KI“ verbreitet sich in nahezu allen Bereichen von Wissensarbeit in bislang nicht gekanntem Maße und geht mit zentralen Veränderungen der dortigen Arbeitsweisen einher. Betriebliche Lösungen für den Einsatz generativer KI fehlen oftmals.
Durch die Nutzung von genKI entstehen neue Gestaltungsbedarfe und Interessenkonflikte welche die betriebliche Gestaltung und Steuerung (Pfeiffer, Nicklich & Schreyer, 2024) bzw. die „digital governance“ (Sciuk, Engert, Gierlich-Joas & Hess, 2025) herausfordern.
Fragestellung: Der Beitrag zeigt auf Basis erster vorliegender empirischer Studien, welche Arbeitswirkungen und Interessenkonflikte sich aufgrund der Nutzung generativer „Alltags-KI“ am Arbeitsplatz in Betrieben entwickeln und diskutiert die möglichen Konsequenzen für die betriebliche Governance, Arbeitsgestaltung und Mitbestimmung. Damit wird die Aushandlung der Nutzung von genKI ins Blickfeld gerückt.

Konzeptioneller Rahmen: Die Untersuchung der Techniknutzung bei der Arbeit erfolgt traditionell in transdisziplinärer Perspektive, idealtypisch beim soziotechnischen Systemansatz des Tavistock Institutes. Demnach bildet eine wechselseitige Optimierung von technischem und sozialem Teilsystem die Voraussetzung für eine effektive Techniknutzung bei der Arbeit.
Dies gilt auch und in verstärktem Maße für Prozesse der Digitalisierung (Pasmore, Winby, Mohrman & Vanasse, 2019). Da die konkrete Technikanwendung immer in soziale Beziehungen eingebettet ist, welche mit unterschiedlichen Interessenlagen und – insbesondere in Betrieben – auch mit ungleichem Machtgefälle einhergeht, ist die Einführung und Nutzung von digitaler Technik Gegenstand von Aushandlungsprozessen (Pfeiffer & Schrape, 2023, S. 136f.). In diese Aushandlungen um die konkrete Nutzungsweise digitaler Technik gehen gesellschaftlich bestimmte Technikerwartungen, betrieblichen Strategien im Kontext der spezifischen Wettbewerbsbedingungen in den jeweiligen Branchen und arbeitspolitischen Leitbildern des Managements ein (Kuhlmann, 2021). Aushandlungsprozesse finden formell und informell auf unterschiedlichen betrieblichen Ebenen statt (Schaupp, 2023). Direkt am Arbeitsplatz bringen Beschäftigte ihre arbeitsbezogenen Interessen auch individuell ein („Job Crafting“ Wrzesniewski & Dutton, 2001), z.B. in der Art und Weise wie sie die Technik nutzen („Technology Crafting“ Law & Varanasi, 2025). Darüber hinaus werden Fragen der Techniknutzung zwischen Mitarbeitenden und Vorgesetzten im Rahmen von Führung und Arbeitsgestaltung verhandelt – unter Einbezug der jeweiligen Fachabteilungen (IT, Personal, Qualitätsmanagement usw.). Die Überwachung und Steuerung des Einsatzes digitaler Technik erfolgt also keineswegs nur auf der Unternehmensebene oder in der Arena der betrieblichen Mitbestimmung.

Auch aus Management-Perspektive wird die Einbindung der Belegschaft als notwendiges Mittel gesehen, um Unterstützung für die digitale Transformation zu erhalten (Sciuk et al., 2025).
In empirischen Projekten hat sich die Nutzung digitaler Technik als *„multidimensional, komplex und konfliktträchtig“* (Rüb, 2021, S. 334) erwiesen. Unser Beitrag geht von einer empirisch fundierten *„Konfliktlandkarte“* bei der Digitalisierung der Arbeit (Rüb, 2021) aus. Diese beschreibt typische interessenpolitische Aushandlungsfelder bei der Einführung und Nutzung digitaler Technik und ihre Wirkungen. So wirft das Konfliktfeld „Technisierung“, weil sie Tätigkeiten und Qualifikationsanforderungen verändert, z. B. die Fragen auf wie die Technik angemessen eingeführt, die Arbeit neu gestaltet und mit Qualifikationswirkungen (Auf- bzw. Abwertungen) umgegangen werden soll. Weitere Aushandlungsfelder sind die Automatisierung, die Datafizierung, die Virtualisierung und die Agilisierung (Rüb, 2021).

Methodisch geht der Beitrag von einer Auswertung von 61 empirischen Studien zu Effekten der Nutzung von genKI am Arbeitsplatz aus. In einer systematischen Analyse werden empirische Belege für die vom Konzept der Konfliktlandkarte postulierten Digitalisierungskonflikte und ihren Wirkungen identifiziert. Die Konfliktlandkarte ist vor der Einführung generativer KI-Anwendungen erstellt worden, berücksichtigt ihr neuartigen Charakteristika also nicht: Aufgrund ihrer Datenbasis sowie ihrer Fähigkeit dialogisch, vermeintlich partnerschaftlich zu kommunizieren und neue, nicht planbare Inhalte generieren zu können, erzeugt ihr Einsatz neuartige Konflikte. Diese wurden identifiziert und systematisiert. Auf dieser Grundlage wird die ursprüngliche „Konfliktlandkarte der Digitalisierung" zu einer „Konfliktlandkarte der Nutzung generativer KI" spezifiziert. In einem dritten Schritt werden dann die Konsequenzen herausgearbeitet, die diese veränderte Konfliktlandkarte für die Aushandlung der Nutzung bedeuten kann. Dabei werden drei zentrale Aushandlungsarenen (Arbeitsgestaltung, digital Governance, Mitbestimmung) betrachtet, um die bisher oftmals getrennte Perspektiven auf Arbeitsgestaltung (i. S. Industrial Engineering) und Mitbestimmung zu integrieren.

Zentrale Ergebnisse: Der Stand der Forschung zur empirischen Nutzung von genKI weist eine Forschungslücke auf: Betriebliche Aushandlungsprozesse und vor allem die Aushandlung im Rahmen der Mitbestimmung werden bislang kaum untersucht. Aus der Empirie können dennoch Konfliktfelder abgeleitet werden, es zeigen sich auch neuartigen arbeitsbezogenen Konflikte, so dass eine deutlich veränderte Konfliktlandkarte entsteht. Zwar resultieren wie auch bei bisheriger Digitalisierung Konflikte aus der Technisierung, der Automatisierung und der Datafizierung, jedoch sind diese teilweise spezifisch „genKI-grundiert" (z. B. Substitution von jetzt auch Nicht-Routine-Tätigkeiten hochqualifizierter Kopfarbeit). Die Konfliktfelder Virtualisierung und Agilisierung verlieren ihre Bedeutung. Sta ttdessen werden die neuen Konfliktfelder „kognitive Entgrenzung" und „ethische Fragen" relevant. Ersteres beschreibt die schwierige Rollenabgrenzung des Menschen gegenüber einer genKI, die mittels einer dialogischen, menschenähnlichen Kommunikation Interaktion simuliert und bei agentischen Systemen auch Handlungsträgerschaft übernimmt (Naik, Toombs, Snellinger, Saponas & Hall, 2025). Ethische Konflikte resultieren u.a. aus Compliance-Risiken, die aus der Fehlerhaftigkeit von genKI-Ergebnissen resultieren („Halluzinationen"), aus fehlender Transparenz der Datengrundlagen und den Fragen von Verantwortung und Haftung.

Auf dieser Grundlage lassen sich erste Thesen zu den neuen Herausforderungen für die Gestaltung und Steuerung der Nutzung von genKI in Organisationen formulieren:

1. Das durch den hohen Anteil an informeller Nutzung von genKI am Arbeitsplatz erzeugte „Regelungsvakuum" kann von Unternehmen mit einem Ansatz der „digital Governance" beantwortet werden (z. B. Anthuvan, Prabhuram, Raju, Maheshwari & Mishra, 2025). Dabei sind jedoch eine technisch-administrativen Sicht auf IT (Carroll, Hassan, Junglas, Hess & Morgan, 2023) und auch ein Verständnis von Top-down-Management (Sciuk et al., 2025) durch ein Konzept von „digital Governance" zu ersetzen, das die Techniknutzung als sozio-technisches System auf Basis von diskursiven Aushandlungen versteht und soziale Praktiken bottom-up zu integrieren lernt.
2. Da Digitalisierungsprozesse wie die genKI-Nutzung keinen abschließenden Charakter haben, sind dauerhafte betriebliche Strukturen zu schaffen, die drei Ebenen der Steuerung integrieren (Bankins, Ocampo, Marrone, Restubog & Woo, 2023; Pfeiffer et al., 2024). Die Arbeit mit genKI am Arbeitsplatz ist professionell zu gestalten („Arbeitsgestaltung"), die Prozesse der digitalen Transformation betrieblich zu steuern („Digital Governance") und ggf. im Rahmen der Mitbestimmung mit Betriebs-/Personalräten zu regeln (Mitbestimmung).
3. Die Herausforderung besteht darin die verschiedenen Initiativen übergreifend zu integrieren und immer wieder einen Ausgleich im Sinne einer „balanced optimization" (Pasmore et al., 2019) zwischen den verschiedenen Ebenen und Arenen herbeizuführen.
4. Im Rahmen der Mitbestimmung sind bestimmte Themen und Vorgehensweisen für einen kollektiven Interessenausgleich bei Technisierungs- und Automatisierungsprozessen etabliert, wie z. B. Vereinbarungen zum Datenschutz, zum Ausschluss von Verhaltenskontrolle und Überwachung, zur

Qualifizierung, Beschäftigungssicherung. Die neuen Konfliktfelder werden jedoch nicht völlig von den bisherigen eingespielten Gestaltungs- und Regelungsmechanismen abgedeckt. Weitere Herausforderung für betriebliche Regelungen stellt darüber hinaus der schnelle Wandel der genKI-Nutzung dar, flexible Anpassungen zu ermöglichen (z. B. mit lernenden Betriebsvereinbarungen) und prozedurale Lösungsstrategien (Matuschek & Kleemann, 2018). Ebenso betrifft die Veränderung verstärkt hochqualifizierte Angestellte, welche bislang weniger im Fokus von Automatisierungsprozessen standen und sich traditionell durch eine stark individualistische Interessenvertretung auszeichneten (Kämpf & Lühr, 2024).

5. Nicht zuletzt muss die genKI-Nutzung im Betrieb als ein iteratives Lernfeld verstanden werden und als Chance, die verschiedenen Interessenlagen im Betrieb immer wieder neu zusammen zu bringen. Erfolgreiche betriebliche Regelungen entstehen durch partizipative Experimentierzyklen (z. B. Pilotprojekte oder Feedback-Loops) und institutionalisierte Reflexionsforen (z. B. KI-Beirat). Mitbestimmung und Arbeitsgestaltung können einander ergänzen, indem die gesammelten Erkenntnisse aus Experimentphasen in verbindliche betriebliche Regelungen einfließen, sowie umgekehrt die Mitbestimmungsbeschlüsse einen breiteren und praxisnahen Bezug zu den tatsächlichen Arbeitsabläufen und Beschäftigtenwie Betriebsinteressen bekommen.

Alles in allem unterstützt der Beitrag mit der empirisch fundierten Konfliktlandkarte für genKI die interdisziplinäre Diskussion über die Aushandlung der Nutzung und der „digital governance“ in Unternehmen. Es deutet sich ein interdisziplinärer Konsens an, dass eine erfolgversprechende Nutzung auf der Grundlage eines sozio-technischen Systemverständnisses und einer partizipativen Gestaltung und Steuerung der Nutzung gelingen könnte.

Literatur:

Anthuvan, T., Prabhuram, D. S., Raju, G., Maheshwari, D. K. & Mishra, A. (2025). Bring Your Own Ai (BYOAI) in the Workplace Patterns, Pitfalls, and the BYOAI-GovTM Framework for Behavioraware Governance. *LEX LOCALIS*, *23*. https://doi.org/10.52152/mmrgh422

Bankins, S., Ocampo, A., Marrone, M., Restubog, S. & Woo, S. (2023). A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. *Journal of Organizational Behavior*. https://doi.org/10.1002/job.2735

Carroll, N., Hassan, N. R., Junglas, I., Hess, T. & Morgan, L. (2023). Transform or be transformed: the importance of research on managing and sustaining digital transformations. *European Journal of Information Systems*, *32*(3), 347–353. https://doi.org/10.1080/0960085X.2023.2187033

Kämpf, T. & Lühr, T. (2024). Angestellte und Mitbestimmung in der digitalen Transformation. *WSI-Mitteilungen*, *77*(1), 34–41. https://doi.org/10.5771/0342-300X-2024-1-34

Kuhlmann, M. (2021). Arbeit in der Digitalisierung: strukturierte Vielfalt der Nutzungsformen und Arbeitswirkungen. In K.-P. Buss, M. Kuhlmann, M. Weißmann, H. Wolf & B. Apitzsch (Hrsg.), *Digitalisierung und Arbeit. Triebkräfte - Arbeitsfolgen - Regulierung* (S. 173–208). Frankfurt; New York: Campus.

Law, M. & Varanasi, R. (2025). Generative AI & Changing Work: Systematic Review of Practitioner-led Work Transformations through the Lens of Job Crafting. *ArXiv*, *abs/2502.08854*. https://doi.org/10.1007/978-3-031-92823-9_10

Matuschek, I. & Kleemann, F. (2018). „Was man nicht kennt, kann man nicht regeln“ Betriebsvereinbarungen als Instrument der arbeitspolitischen Regulierung von Industrie 4.0 und Digitalisierung. *WSI-Mitteilungen*, *71*(3), 227–234. https://doi.org/10.5771/0342-300X-2018-3-227

Naik, S., Toombs, A., Snellinger, A., Saponas, S. & Hall, A. (2025). Exploring Human-AI Collaboration Using Mental Models of Early Adopters of Multi-Agent Generative AI Tools. arXiv. https://doi.org/10.48550/arxiv.2510.06224

Pasmore, W., Winby, S., Mohrman, S. A. & Vanasse, R. (2019). Reflections: Sociotechnical Systems Design and Organization Change. *Journal of Change Management*, *19*(2), 67–85. https://doi.org/10.1080/14697017.2018.1553761
Pfeiffer, S. (2025). (Generative) Künstliche Intelligenz (KI) als Kollegin? Gestaltung und Mitbestimmung aus Sicht der Beschäftigten. In M. Wannöffel, F. Hoose, C. Niewerth & H.-J. Urban (Hrsg.), *Mitbestimmung und Partizipation 2030: Demokratische Perspektiven auf Arbeit und Beschäftigung* (S. 247–264). Baden-Baden: Nomos Verlagsgesellschaft GmbH & Co. KG. https://doi.org/10.5771/9783748953081
Pfeiffer, S., Nicklich, M. & Schreyer, J. (2024). Digitalisierung der Arbeitswelten – die systemische Transformation verstehen. Ein Ausblick. In S. Pfeiffer, M. Nicklich, M. Henke, M. Heßler, M. Krzywdzinski & I. Schulz-Schaeffer (Hrsg.), *Digitalisierung der Arbeitswelten: Zur Erfassbarkeit einer systemischen Transformation* (S. 559–584). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-44458-7
Rüb, S. (2021). Digitalisierung als Konfliktfeld betrieblicher Arbeitsbeziehungen. In K.-P. Buss, M. Kuhlmann, M. Weißmann, H. Wolf & B. Apitzsch (Hrsg.), *Digitalisierung und Arbeit. Triebkräfte - Arbeitsfolgen - Regulierung*. Frankfurt; New York: Campus.
Schaupp, S. (2023). Was ist Technopolitik? Aushandlungsarenen in der digitalisierten Arbeitswelt. In T. Carstensen, S. Schaupp & S. Sevignani (Hrsg.), *Theorien des digitalen Kapitalismus. Arbeit, Ökonomie, Politik und Subjekt* (S. 326–343). Berlin: Suhrkamp.
Sciuk, C., Engert, S., Gierlich-Joas, M. & Hess, T. (2025). How Companies Navigate the (Un)Charted Waters of Digital Transformation. *California Management Review*, *68*(1), 5–31. https://doi.org/10.1177/00081256251377811
Wrzesniewski, A. & Dutton, J. E. (2001). Crafting a Job: Revisioning Employees as Active Crafters of Their Work. *The Academy of Management Review*, *26*(2), 179. https://doi.org/10.2307/259118

Governance digitaler Transformation zwischen Technologie und Handlungsfähigkeit: Eine Fallstudie zur Implementierung von Transformationskompetenz in einem Maschinenbauunternehmen

Kendra Pöhlmann | OTH Regensburg; Ruth Nürnberger | Brückner Maschinenbau

1. Einleitung

Digitale Transformation gilt in Wirtschaft und Arbeitswelt zunehmend als zentrale Voraussetzung für Wettbewerbsfähigkeit, Innovationsfähigkeit und Zukunftssicherung (Chanias et al.2019). Gleichzeitig wird immer deutlicher, dass digitale Transformation nicht auf technologische Modernisierung reduziert werden kann, sondern tiefgreifende organisationale, kulturelle und relationale Veränderungsprozesse umfasst (Vial 2021). Organisationen stehen dabei unter erheblichem Transformationsdruck, der nicht nur durch technologische Entwicklungen wie Künstliche Intelligenz, datenbasierte Steuerung oder ERP-Transformationen entsteht, sondern auch durch globale Unsicherheiten, volatile Märkte, Nachhaltigkeitsanforderungen und veränderte Formen der Zusammenarbeit.
Digitale Transformation ist dabei kein linearer oder eindeutig steuerbarer Prozess, sondern vielmehr ein dynamisches und widersprüchliches „moving target“ (Pfeiffer et al.2024). Organisationen sehen sich mit konkurrierenden Rationalitäten, widersprüchlichen Zielsetzungen und zunehmender Komplexität konfrontiert. Vor diesem Hintergrund gewinnt die Frage an Bedeutung, wie digitale Transformation unter Bedingungen von Unsicherheit, Dynamik und organisationalen Spannungen gestaltet und koordiniert werden kann (Carroll et al.2023; Sciuk et al.2025).

Der vorliegende Beitrag greift diese Diskussion auf und untersucht digitale Transformation aus einer Governance-Perspektive. Im Zentrum steht dabei die These, dass digitale Transformation nicht primär an

fehlender Technologie oder ungeeigneten Methoden scheitert, sondern an begrenzter individueller und kollektiver Handlungsfähigkeit im Umgang mit Ambiguität, Unsicherheit und organisationalen Spannungen.
Zur theoretischen Rahmung wird das Konzept der Transformationkompetenz herangezogen. Transformationkompetenz beschreibt ein organisations- und entwicklungspsychologisch fundiertes Kompetenzbündel, das u. a. emotionale Selbstregulation, Reflexionsfähigkeit, Ambiguitätstoleranz, kollaborative Sinnkonstruktion und die Fähigkeit zum Perspektivwechsel umfasst. Der Beitrag argumentiert, dass diese Kompetenzen eine bislang unterbelichtete Voraussetzung reflexiver Governance digitaler Transformation darstellen.
Empirisch basiert der Beitrag auf einer qualitativen Fallstudie aus einem international tätigen Maschinenbauunternehmen, das sich aktuell in mehreren parallelen Transformationsprozessen befindet, darunter insbesondere die Einführung von SAP S/4HANA im Greenfield-Ansatz.

2. Governance digitaler Transformation als Mehrebenenphänomen

Governance digitaler Transformation wird im vorliegenden Beitrag nicht ausschließlich als formales Steuerungs- oder Managementproblem verstanden, sondern als komplexes Zusammenspiel organisationaler Entscheidungslogiken, informeller Praktiken, Koordinationsmechanismen und sozio-technischer Arrangements (Mayntz 2004; Wilkin & Chenhall 2010). Damit schließt der Beitrag an Governance-Verständnisse an, die digitale Transformation als relationalen und multiperspektivischen Aushandlungsprozess begreifen (Fox & Ward 2008; Campos-Castillo et al 2025). Managementpraktiken, Projektlogiken und digitale Technologien strukturieren Transformation zwar wesentlich mit, reichen jedoch nicht aus, um organisationale Veränderungsprozesse zu erklären oder nachhaltig zu stabilisieren.
Insbesondere groß angelegte digitale Transformationsprojekte erzeugen organisationale Spannungen, die sich nicht allein über technische Implementierungslogiken bearbeiten lassen. Vielmehr entstehen Widersprüche zwischen Stabilität und Flexibilität, Standardisierung und Autonomie, Effizienz und Reflexion oder technologischer Beschleunigung und organisationaler Verarbeitungskapazität. Pfeiffer et al.(2024) beschreiben digitale Transformation entsprechend als systemische Transformation mit gegenläufigen Dynamiken und widersprüchlichen Anforderungen. Auch Sciuk et al.(2025) betonen, dass Unternehmen digitale Transformation zunehmend unter Bedingungen permanenter Unsicherheit und fehlender Zielklarheit navigieren.
Die Fallstudie greift diese Spannungen auf und betrachtet digitale Transformation daher explizit als Mehrebenenphänomen, das sich gleichzeitig auf globaler Umwelt- und Marktebene, auf organisationaler Ebene, innerhalb von Teams und Kooperationsbeziehungen sowie auf Ebene individueller Wahrnehmungs-, Reflexions- und Handlungsmuster vollzieht. Damit verbindet der Beitrag organisationale Governance-Perspektiven mit organisationspsychologischen und kompetenzorientierten Zugängen.

3. Fallkontext: Transformation im Maschinen- und Anlagenbau

Die empirische Grundlage des Beitrags bildet eine qualitative Fallstudie bei Brückner Maschinenbau, einem international tätigen familiengeführten Maschinenbauunternehmen mit rund 1.200 Mitarbeitenden und globaler Marktpräsenz.
Das Unternehmen befindet sich aktuell in mehreren überlagerten Transformationsprozessen. Dazu zählen unter anderem die stärkere Integration in eine gruppenweite Matrixstruktur, die Etablierung neuer Business Units, die Integration weiterer Unternehmensbereiche sowie die Einführung von SAP S/4HANA im Greenfield-Ansatz.
Die SAP-Transformation umfasst die Neugestaltung von End-to-End- und Intercompany-Prozessen sowie den Aufbau einer gruppenweit vereinheitlichten IT-Infrastruktur. Gleichzeitig greifen diese Veränderungen tief in organisationale Routinen, Entscheidungswege, Rollenverständnisse und Formen der Zusammenarbeit ein.
Die Fallstudie verdeutlicht damit exemplarisch, dass digitale Transformation nicht nur technologische Modernisierung bedeutet, sondern organisationale Identitäten, Kooperationsformen und Macht- bzw. Koordinationsstrukturen verändert.

Besonders deutlich zeigt sich dies in den Spannungen zwischen klassischen projektlogischen Steuerungsansätzen und den Anforderungen adaptiver, lernorientierter Organisationsentwicklung. Während das Kerngeschäft des Maschinen- und Anlagenbaus weiterhin stark von langfristiger Planung, technischer Exzellenz und hoher Prozesssicherheit geprägt ist, erzeugen volatile Märkte, globale Lieferkettenprobleme und digitale Transformationsdynamiken gleichzeitig einen erheblichen Bedarf an Flexibilität, Reflexionsfähigkeit und bereichsübergreifender Zusammenarbeit.
Die Organisation bewegt sich damit in einem Spannungsfeld konkurrierender Rationalitäten und permanenter Aushandlungsprozesse.

4. Transformationkompetenz als reflexive Governance-Praxis
Zur Analyse dieser Dynamiken wird im Beitrag das Konzept der Transformationkompetenz herangezogen. Transformationkompetenz versteht Transformation nicht primär als technische oder strukturelle Aufgabe, sondern als Fähigkeit von Individuen, Teams und Organisationen, mit Unsicherheit, Widersprüchen und komplexen Veränderungsdynamiken konstruktiv umzugehen.
Der Ansatz integriert Perspektiven aus organisations- und entwicklungspsychologischen Ansätzen, Forschung zu Future Skills (Ehlers 2020), Bildung für nachhaltige Entwicklung (Vare et al.2019; Brundiers et al.2021) sowie transformationstheoretischen und organisationsbezogenen Diskursen. Gleichzeitig bestehen deutliche Anschlussstellen zu aktuellen Debatten um die Inner Development Goals und die Bedeutung innerer Entwicklung für gesellschaftliche und organisationale Transformationsprozesse (Ankrah et al.2023; Wamsler et al.2024).

Im organisationalen Kontext umfasst Transformationkompetenz insbesondere emotionale Selbstregulation, Reflexionsfähigkeit, Ambiguitätstoleranz, kollaborative Sinnkonstruktion, Perspektivintegration sowie die Fähigkeit zur kooperativen Gestaltung organisationaler Veränderung. Damit knüpft der Beitrag an organisationspsychologische Arbeiten zu Selbstregulation (Baumeister 2004), Unsicherheit und organisationaler Veränderung (Ameln & Iwers 2021) sowie an Diskussionen um die Bedeutung „weicher“ Faktoren in der digitalen Transformation an (Nink 2020).
Die Fallstudie zeigt, dass insbesondere informelle Praktiken und organisationsinterne Übersetzungsleistungen eine zentrale Rolle spielen. Transformation wurde nicht allein über klassische Projektmanagementlogiken bearbeitet, sondern zusätzlich über Reflexionsräume, Leadership Circles, Mentoring- und Buddy-Programme, Coaching- und Sparring-Formate sowie organisationsinterne Change- und Kommunikationspraktiken. Diese Formate dienten weniger der direkten Steuerung von Transformation als vielmehr der Stabilisierung kollektiver Handlungsfähigkeit unter Bedingungen organisationaler Unsicherheit.
Besondere Bedeutung kommt dabei sogenannten Boundary-Spanning-Rollen zu. Organizational Change Management fungierte im Rahmen des SAP-Projekts als vermittelnde Instanz zwischen technologischer Projektlogik, organisationalen Routinen und kulturellen Dynamiken. Boundary Spanner übersetzen dabei nicht nur Informationen, sondern moderieren Spannungen zwischen unterschiedlichen organisationalen Logiken und tragen zur Synchronisation organisationaler Veränderungsprozesse bei.
Damit schließt der Beitrag an Boundary-Spanning-Forschung an, die Boundary Spanner als zentrale Akteure komplexer Governance- und Koordinationsprozesse beschreibt (Williams 2013; van Meerkerk & Edelenbos 2014). Auch Ferguson (2005) zeigt die Bedeutung relationaler Nähe und vertrauensbasierter Vermittlungsleistungen für erfolgreiche Koordinationsprozesse. Im Kontext digitaler Transformation gewinnen solche Rollen zusätzlich an Bedeutung, da technologische, organisationale und kulturelle Veränderungsdynamiken zunehmend miteinander verschränkt sind.
Die Fallstudie verweist damit insgesamt auf die Relevanz translationaler, reflexiver und lernorientierter Governance-Praktiken digitaler Transformation.

5. Diskussion: Governance als Handlungsfähigkeitsproblem

Die Fallstudie legt nahe, Governance digitaler Transformation stärker als Handlungsfähigkeits- und Koordinationsproblem zu verstehen. Digitale Technologien, Managementpraktiken und Projektstrukturen bleiben wichtige Bestandteile organisationaler Governance. Gleichzeitig zeigt die Analyse, dass digitale Transformation wesentlich von der Fähigkeit abhängt, organisationale Spannungen reflexiv zu bearbeiten und kollektive Handlungsfähigkeit unter Bedingungen von Unsicherheit aufrechtzuerhalten. Damit erweitert der Beitrag bestehende Governance-Diskurse um eine organisationspsychologische und kompetenzorientierte Perspektive. Besonders relevant erscheint dabei die Beobachtung, dass viele organisationale Herausforderungen nicht primär auf fehlende Technologien oder mangelndes Wissen zurückzuführen sind, sondern auf begrenzte organisationale Verarbeitungskapazitäten. Reagieren statt Vorsorgen, Ad-hoc-Maßnahmen statt strategischer Reflexion sowie Fragmentierung zwischen organisationalen Funktionen stellen wiederkehrende Muster dar.

Die Ergebnisse stützen damit Ansätze reflexiver Governance, die digitale Transformation nicht ausschließlich über Steuerung und Kontrolle, sondern stärker über Lern-, Übersetzungs- und Aushandlungsprozesse begreifen. Zugleich verdeutlicht die Fallstudie die Bedeutung von Kultur als verbindendes Element zwischen Prozessen, Strukturen und individuellem Handeln. Unternehmenskultur fungiert dabei weniger als stabiler Hintergrundfaktor, sondern vielmehr als dynamischer „Kleber“ organisationaler Transformationsprozesse.

6. Fazit

Der Beitrag zeigt, dass Governance digitaler Transformation nicht allein als technisches oder managementbezogenes Steuerungsproblem verstanden werden kann. Vielmehr verweist die Fallstudie auf die Bedeutung reflexiver, lernorientierter und kompetenzbasierter Governance-Praktiken, die individuelle, teambezogene und organisationale Ebenen miteinander verbinden. Transformationkompetenz fungiert dabei als verbindendes Element zwischen technologischer Transformation, organisationaler Koordination und menschlicher Handlungsfähigkeit. Die Ergebnisse legen nahe, digitale Transformation stärker als sozio-technischen und kulturellen Aushandlungsprozess zu betrachten, in dem Boundary-Spanning-Praktiken, Reflexionsräume und kollaborative Lernprozesse zentrale Voraussetzungen organisationaler Veränderungsfähigkeit darstellen.

Literatur

Ameln, F. von, & Iwers, T. (2021). Unsicherheit und Ungewissheit. Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie, 52(4), 563–566.

Ankrah, D., Bristow, J., Hires, D., & Henriksson, J. (2023). Inner Development Goals: from inner growth to outer change. Special Issue 25, 82–87.

Baumeister, R. F. (Hrsg.) (2004). Handbook of self-regulation: Research, theory, and applications. Guilford Press.

Brundiers, K., Barth, M., Cebrián, G., Cohen, M., Diaz, L., Doucette-Remington, S. et al. (2021). Key competencies in sustainability in higher education—toward an agreed-upon reference framework. Sustainability Science, 16(1), 13–29.

Ehlers, U.-D. (2020). Future Skills. Springer Fachmedien Wiesbaden.

Ferguson, R. J. (2005). Contractual governance, relational governance, and the performance of interfirm service exchanges: The influence of boundary-spanner closeness. Journal of the Academy of Marketing Science, 33(2), 217–234.

Aguilera, R. V., Filatotchev, I., Gospel, H., & Jackson, G. (2008). An organizational approach to comparative corporate governance: Costs, contingencies, and complementarities. Organization Science, 19(3), 475–492.

Campos-Castillo, C., Freedman, J., Kang, X., Atkinson, E., Galinkala, P., Giri, D. et al.(2025). Humanizing digital governance: Human-centered sociological scholarship into digital governance. Information, Communication & Society, 1–6.

Carroll, N., Hassan, N. R., Junglas, I., Morgan, L., & Hess, T. (2023). Transform or be transformed: The importance of research on managing and sustaining digital transformations. European Journal of Information Systems, 32(3), 347–353.
Chanias, S., Myers, M. D., & Hess, T. (2019). Digital transformation strategy making in pre-digital organizations: The case of a financial services provider. Journal of Strategic Information Systems, 28(1), 17–33.
Fox, N. J., & Ward, K. J. (2008). What governs governance, and how does it evolve? The sociology of governance-in-action. British Journal of Sociology, 59(3), 519–538.
Mayntz, R. (2004). Governance Theory als fortentwickelte Steuerungstheorie? MPIfG Working Paper 04/1.
Pfeiffer, S., Nicklich, M., & Schreyer, J. (2024). Digitalisierung der Arbeitswelten – die systemische Transformation verstehen. In S. Pfeiffer et al. (Hrsg.), Digitalisierung der Arbeitswelten (S. 559–584). Springer VS.
Sciuk, C., Engert, S., Gierlich-Joas, M., & Hess, T. (2025). How companies navigate the (un)charted waters of digital transformation. California Management Review, 68(1), 5–31.
Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The sociotechnical axis of cohesion for the IS discipline: Its historical legacy and continued relevance. MIS Quarterly, 43(3), 695–719.
Vial, G. (2021). Understanding digital transformation: A review and a research agenda. In A. Hinterhuber, T.Vescovi & F. Checchinato (Hrsg.), Managing Digital Transformation (S. 13–66). Routledge.
Wilkin, C. L., & Chenhall, R. H. (2010). A review of IT governance: A taxonomy to inform accounting information systems. Journal of Information Systems, 24(2), 107–146.
Williamson, O. E. (1996). The mechanisms of governance. Oxford University Press.

COMMUNICATION & SOCIETY Session 1
Digital Public Sphere: Conditions and Framework for a Constructive Debating Culture

19.11.2026, 10:15-12:15

Language: English

Programme Committee: Andreas Wenninger | bidt; Hannah Schmid-Petri | bidt & University of Passau; Ute Schmid | bidt & University of Bamberg; Andreas Jungherr | bidt & University of Bamberg

Session Chair: Hannah Schmid-Petri | bidt & University of Passau

AI-supported stakeholder simulation as a tool for digital participation: The Eckernkamp Compass

Nicholas H. Müller | SHIFT; Michael Müßig | SHIFT; Markus Oermann | SHIFT; Marie Herz | SHIFT; Hakan Arda | SHIFT; Julian Sehne | SHIFT

1. Research Question and Relevance

The digitization of government and social services is one of the key challenges facing modern societies. Concepts and strategies often refer to human-centered or civic-centered design, that is, the need to place citizens' perspectives at the center from the very beginning. In practice, however, there is a significant gap: Mergel et al. (2019) demonstrate in a comprehensive expert study that citizen participation remains underrepresented in digital transformation processes in the public sector, with process optimization taking precedence over genuine user-centeredness. The result is digital services that fail to address people's actual needs and lived realities. The Eckernkamp Compass addresses this gap between aspiration and reality: It begins at the point where citizens' perspectives must be systematically made visible – early enough to still influence development processes. The Eckernkamp Compass is an AI-supported research and development project of the SHIFT Institute that uses Silicon Sampling and a RAG architecture to make societal stakeholder perspectives visible at an early stage. To achieve this, scientifically validated research data is fed into an AI system that simulates relevant citizen and stakeholdergroups as interactive avatars, thereby specifically enhancing real-world participatory processes.

The SHIFT Institute for the Digital Society has made it its mission to provide evidence-base impetus for shaping digital transformation in a responsible manner that prioritizes the common good. The Eckernkamp Compass, the institute's flagship project, puts this mission into practice. This article asks: Can generative AI be used as a tool to better take citizens' interests into account? More specifically, we examine the methodological and institutional conditions under which AI-based stakeholder simulations can make digitalization processes more inclusive and of higher quality.

2. Theoretical and Conceptual Framework

Deliberative democracy and digital participation: Citizens' assemblies, citizens' juries, and consensus conferences are proven participatory formats. The OECD (2020) documents that they strengthen democracy, foster trust, and enable better political decisions. Tessler et al. (2024) further demonstrate empirically that AI can help find common ground in such processes. This suggests that generative AI should be understood as a tool for more constructive participation.

Silicon Sampling and its limitations: Argyle et al. (2023) show that LLMs can simulate human response patterns under certain conditions. However, recent studies reveal systematic weaknesses: even small changes in wording lead to significant discrepancies between AI and human responses, and different models react very differently to new questions (Schröder et al., 2025).

Sarstedt et al. (2024) draw a clear conclusion: Silicon Sampling is particularly suitable for early phases of the research process, such as qualitative preliminary studies and pretests, but not as a complete replacement for human interviews. Majumdar et al. (2026) further demonstrate that collective AI methods can be robust against model biases under certain conditions. Wihbey and D'Alonzo (2025) recommend LLMs as an effective complement to interviews with individuals.
The SHIFT Institute follows this nuanced approach: using silicon sampling to complement realworld participation. In addition, the language models with RAG extensions are iteratively improved and validated at the institute to ensure representative results.
AI Transparency and Trust: In a representative survey conducted in Germany, Jungherr and Rauchfleisch (2025) demonstrated the existence of an "AI penalty": People are skeptical when AI is used in participatory processes. This can create new inequalities. The implication for design: Transparency is mandatory. Lewis et al. (2021) developed the RAG architecture, in which model responses are linked to external source documents, enabling traceability.

3. Methodological Approach: System Architecture of the Eckernkamp Compass
Construction of the Knowledge Corpus: Scientific studies, including data from open repositories such as Zenodo – which can also be used for citizen science projects – are automatically fed into the system. The system breaks down the texts into semantic units, indexes them in a vector database, and thus enables precise semantic similarity searches (Lewis et al., 2021; Ma et al., 2026; Monir et al., 2024). Qualitative stakeholder interviews and quantitative studies form the knowledge base.
Semantic Search and Answer Generation: When someone asks a question, the system identifiesthe most relevant text passages using semantic similarity search. These are passed to the language model as context. Based on this, the model generates a verifiable answer in real time. Researchers, and in the future, all users, can trace which sources influenced the answer.
Immersive output and channel diversity: The stakeholder agents appear as photorealistic avatars, rendered using Unreal Engine 5 and MetaHuman (see Figure 1) (Zimmer et al., 2026). They can be experienced via browsers, mobile devices, and, in collaboration with the MAVEL Lab at THWS1, via immersive cave systems (see Figure 2). In the prototype, Convai handles the control of speech, animation, and knowledge access (Benkhoui, 2024; Mukherjee, 2023). In the long term, a fully locally operated system is planned – based on RAGFlow and Ollama (InfiniFlow, 2023/2026). This ensures that control over sensitive input data remains with the institution.



Figure 1: Meta-Human Avatar



Figure 2: immersive cave system (3D)

Iterative validation: AI can support human participation and also incorporate the voices of those who otherwise would not or could not participate (Majumdar et al., 2026). The SHIFT approach: We use AI in a transparent, methodologically sound manner, with consistent, iterative validation, to ensure responsible outcomes (Wihbey & D'Alonzo, 2025).

4. Key Theses of this Presentation

The following theses of this paper arise from the theoretical foundations outlined above:

Thesis 1: Constructive participation requires preparation: Many participatory formats in digitalization processes, as well as corresponding civic-centered design approaches, fail in practice because relevant perspectives are identified too late or the effort required for sustained stakeholder participation is too great. AI-based simulation can reduce information asymmetries early on. It prepares the ground for real-world participatory processes and complements them in a resource-efficient manner.

Thesis 2: Transparency is a prerequisite for legitimacy: People will only accept AI in participatory processes if sources and limitations are disclosed. A RAG architecture with traceable source references is therefore not a technical gimmick, but a necessity.

Thesis 3: Data sovereignty as the basis of trust: A locally operated system (RAG-Flow + Ollama) ensures institutional control over sensitive participation data. This addresses the central concerns of accountability and integrity in digital systems.

5. Interdisciplinary Interoperability

This presentation combines computer science (RAG, vector databases, immersive systems, human- computer interaction) with insights from governance and participatory research on digitization projects. It addresses the conference's interdisciplinary focus. The presentation demonstrates how algorithms and interfaces can be designed to ensure transparency and human oversight and to achieve civic-centered design. The SHIFT Institute is open to research collaborations and pilot partnerships. The presentation will be delivered in English or German.

Literature

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Benkhoui, Y. (2024, January 8). *Spotlight: Convai Reinvents Non-Playable Character Interactions*. NVIDIA Technical Blog. https://developer.nvidia.com/blog/spotlight-convai-reinvents-non-playable-characterinteractions/

InfiniFlow. (2026). *Infiniflow/ragflow* [Python]. https://github.com/infiniflow/ragflow (Original work published 2023)

Jungherr, A., & Rauchfleisch, A. (2025). *Artificial intelligence in deliberation: The AI penalty and the emergence of a new deliberative divide.* Government Information Quarterly, 42(4), 102079. https://doi.org/10.1016/j.giq.2025.102079

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel,

T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* (arXiv:2005.11401). arXiv. https://doi.org/10.48550/arXiv.2005.11401

Ma, L., Zhang, R., Han, Y., Yu, S., Wang, Z., Ning, Z., Zhang, J., Xu, P., Li, P., Qiao, Z., Ju, W., Chen, C., Wang, D., Liu, K., Wang, P., Wang, P., Fu, Y., Liu, C., Zhou, Y., & Lu, C.-T. (2026). *A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge* (arXiv:2310.11703). arXiv. https://doi.org/10.48550/arXiv.2310.11703

Majumdar, S., Elkind, E., & Pournaras, E. (2026). *Generative AI voting: Fair collective choice is resilient to LLM biases and inconsistencies* (arXiv:2406.11871). arXiv. https://doi.org/10.48550/arXiv.2406.11871

Mergel, I., Edelmann, N., & Haug, N. (2019). Defining digital transformation: Results from expert interviews. *Government Information Quarterly*, *36*(4), 101385. https://doi.org/10.1016/j.giq.2019.06.002

Monir, S. S., Lau, I., Yang, S., & Zhao, D. (2024). *VectorSearch: Enhancing Document Retrieval with Semantic Embeddings and Optimized Search* (arXiv:2409.17383). arXiv. https://doi.org/10.48550/arXiv.2409.17383
Mukherjee, P. (2023). *Introducing Convai: Conversational and Agentic AI Infrastructure for Virtual Worlds*. https://convai.com/blog/introducing-convai
OECD. (2020, June 9). *Innovative Citizen Participation and New Democratic Institutions: Catching the Deliberative Wave*. OECD Publishing. OECD. https://doi.org/10.1787/339306da-en
Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). *Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines.* Psychology & Marketing, 41(6), 1254–1270. https://doi.org/10.1002/mar.21982 Schröder, S., Morgenroth, T., Kuhl, U., Vaquet, V., & Paaßen, B. (2025). Large Language Models Do Not Simulate
Human Psychology (arXiv:2508.06950). arXiv. https://doi.org/10.48550/arXiv.2508.06950
Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., & Summerfield, C. (2024). AI can help humans find common ground in democratic deliberation. Science, 386(6719), eadq2852. https://doi.org/10.1126/science.adq2852
Wihbey, J., & D'Alonzo, S. (2025). AI Simulations of Audience Attitudes and Policy Preferences: "Silicon Sampling" Guidance for Communications Practitioners (SSRN Scholarly Paper No. 5533958). Social Science Research Network. https://doi.org/10.2139/ssrn.5533958
Zimmer, C., Kruse, L., Rolvien, L., Gallinat, J., & Geiger, C. (2026). Transformation of MetaHumans for a Generic XR Workflow. In L. T. De Paolis, P. Arpaia, & M. Sacco (Eds.), Extended Reality (pp. 73–92). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-97778-7_5

Platforms, Messengers and Messages: What Shapes Deliberative Engagement in Climate Scientists' Social Media Debates?

Stephan Schlögl | University of Passau; Paula Kolhep | University of Passau; Hannah Schmid-Petri | University of Passau & bidt

Introduction

Calls for scientists to engage publicly with their research have grown louder, but what constitutes meaningful engagement remains conceptually underspecified (Weingart et al., 2021). Social media provide infrastructure for direct exchange between researchers and broad publics (Cˆot´e & Darling, 2018), yet the conditions that foster high-quality interactions are poorly understood.

We argue that deliberative theory can provide a normative anchor to fill this conceptual gap and test our approach empirically with regard to climate scientists based in Germany. The paper addresses two questions. RQ1: How does deliberative quality differ across platforms? RQ2:
What message-, messenger-, and platform-level factors shape it?

Theory: Deliberativeness in Science Communication

Weingart et al. (2021) document that public engagement with science (PES) scholarship lacks conceptual precision; deliberative theory, with its rich tradition of operationalizable criteria (Friess & Eilders, 2015), can fill this gap. The goals of deliberative democracy: knowledge gain, social trust, political empowerment, map closely onto those attributed to PES (education, legitimation, democratization). Drawing on Friess and Eilders (2015), we adapt six dimensions to the science communication context: rationality (use of arguments and evidence), interactivity (substantive engagement with the preceding message), inclusiveness (absence of jargon and speaker role denial), civility (absence of disrespectful speech toward different actor groups), common good reference (absence of generalized science skepticism or climate change denial), and constructiveness (provision of sources, genuine questions, or solution proposals).

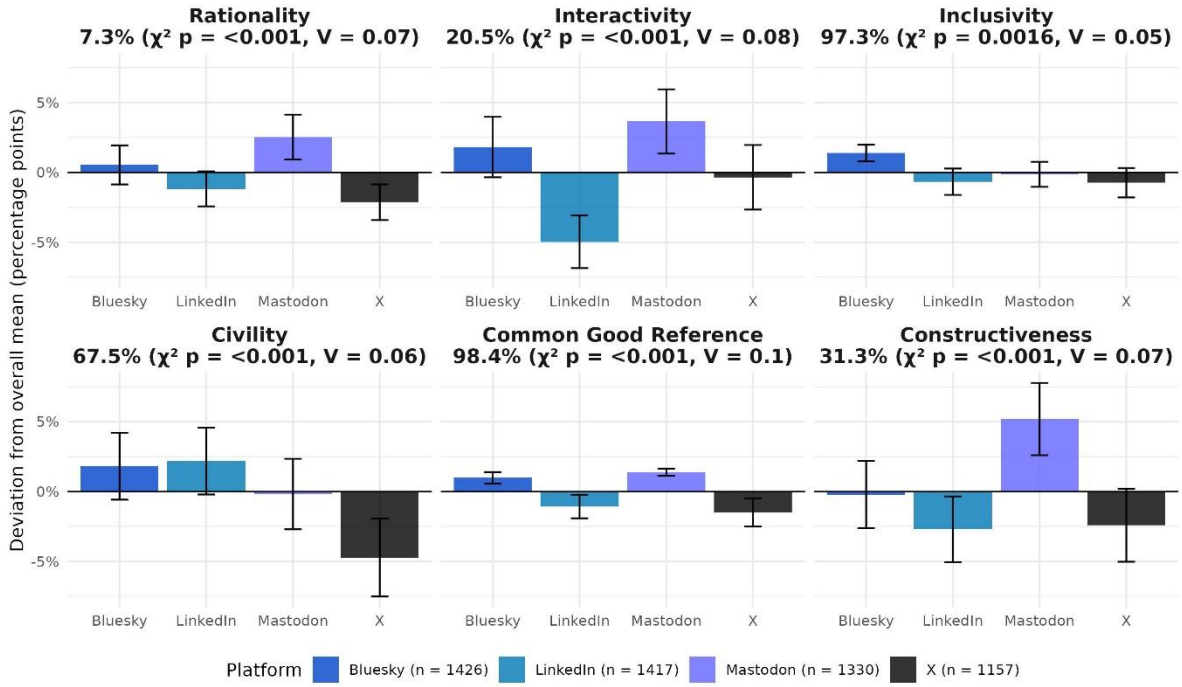
Data and Methods

We identified 1,839 climate scientists based in Germany and located their accounts on Twitter/X, LinkedIn, Mastodon, BlueSky, and Instagram via semi-automated search. Between June and August 2024, up to 500 recent posts per account were collected including the respective comments. A stratified random sample of posts with at least two comments was drawn (95% CI, MoE 5%). Instagram was excluded due to low rates of climate-related content. The final sample comprised N = 701 climate-related posts and N = 5,330 comments. Six coders applied a detailed codebook measuring each deliberativeness dimension dichotomously. Platform differences (RQ1) between comments were examined using chi-square tests and Cram´er's V. For RQ2, we estimated six binary logistic multilevel models (one per dimension) with a random intercept at the researcher level using glmer (Bates et al., 2015) as to predict the deliberative quality of comments. Predictors included an autoregressive term (deliberativeness of the parent message), climate topic indicators, political content, thread depth, whether the parent message was authored by the researcher, researcher gender, log-standardized follower count, and platform dummies.

Results

RQ1: Figure 1 shows deviations from the overall mean by platform and dimension. All deviations fall within roughly five percentage points and Cramér's V is consistently weak, indicating that platform differences, though statistically significant, are modest. Twitter/X scores below average on all six dimensions, most markedly on civility. Mastodon scores highest, particularly on rationality, common good reference, and constructiveness. LinkedIn is notably low on interactivity, likely reflecting the prevalence of congratulatory comments on that professional networking platform.

Figure 1: Deviation from overall mean by platform and deliberativeness dimension (N = 5,330).



RQ2: Platform effects on civility and common good reference persist after controlling for other factors (Table 1). The most consistent finding across all six models is the autoregressive effect.

Table 1: Multilevel logistic regression table for different deliberativeness dimensions (N = 5,330).

	Rationality	Interactivity	Inclusivity	Civility	Common Good Reference	Constructiveness
(Intercept)	−3.124***	−1.276***	1.843***	1.976***	4.621**	−1.254***
	(0.220)	(0.195)	(0.365)	(0.198)	(1.753)	(0.198)
Previous message						
Autoregressor	0.690***	0.611***	1.385***	0.397***	1.999***	0.677***
	(0.137)	(0.107)	(0.225)	(0.091)	(0.468)	(0.071)
Topic: Physical basics	−0.295	0.807***	0.340	0.144	0.326	0.024
	(0.211)	(0.119)	(0.327)	(0.115)	(0.425)	(0.107)
Topic: Mitigation	0.325**	0.097	−0.195	−0.060	−0.211	0.105
	(0.125)	(0.087)	(0.210)	(0.083)	(0.286)	(0.074)
Topic: Adaptation	0.369	0.074	−0.016	−0.120	0.908	−0.134
	(0.238)	(0.189)	(0.455)	(0.177)	(1.042)	(0.165)
Topic: Consequences	0.073	−0.051	−0.664**	−0.338***	−1.137***	0.129
	(0.126)	(0.087)	(0.206)	(0.077)	(0.265)	(0.073)
Topic: Metadebate	0.280	0.172	0.351	−0.022	−0.979*	0.005
	(0.215)	(0.153)	(0.441)	(0.146)	(0.413)	(0.134)
Political issue	0.172	0.046	0.041	0.211***	0.040	−0.092
	(0.100)	(0.073)	(0.178)	(0.061)	(0.204)	(0.060)
Engagement	−0.055	0.005	0.222	−0.058	0.098	−0.000
	(0.066)	(0.046)	(0.124)	(0.046)	(0.148)	(0.039)
Thread depth	0.108***	0.078***	−0.097*	0.267***	−0.046	0.048*
	(0.025)	(0.021)	(0.044)	(0.056)	(0.128)	(0.019)
Author is researcher	−0.359**	−0.953***	0.441*	−2.185***	−0.671	−0.448***
	(0.137)	(0.096)	(0.224)	(0.130)	(0.405)	(0.084)
Researcher						
Followers/contacts (logz)	−0.247***	−0.265***	0.248*	0.047	0.120	−0.249***
	(0.066)	(0.055)	(0.102)	(0.049)	(0.256)	(0.052)
Gender (male = 0)	−0.551*	−0.553*	0.924*	0.378	0.520	−0.732**
	(0.264)	(0.246)	(0.462)	(0.215)	(1.186)	(0.268)
Platforms (X = 0)						
Bluesky	0.161	−0.327*	1.202***	0.389***	2.043***	−0.244*
	(0.192)	(0.130)	(0.350)	(0.117)	(0.460)	(0.113)
Mastodon	0.485**	−0.030	0.183	0.557***	3.237***	0.106
	(0.181)	(0.123)	(0.279)	(0.113)	(0.655)	(0.107)
LinkedIn	0.166	−0.332*	0.068	0.620***	0.679	−0.162
	(0.204)	(0.155)	(0.295)	(0.128)	(0.571)	(0.133)
R2 Marginal	0.092	0.140	0.155	0.375	0.181	0.066
R2 Conditional	0.112	0.195	0.265	0.425	0.683	0.187
ICC	0.022	0.064	0.130	0.080	0.613	0.129

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

The deliberative quality of the parent message is the strongest predictor of the quality of the reply. Rational messages attract rational replies, civil messages attract civil replies, and so on (deliberative reciprocity; cf. Naab et al., 2023). Thread depth shows positive effects on rationality, interactivity, civility, and constructiveness, suggesting deliberative quality improves rather than degrades over extended exchanges (cf. Niemeyer et al., 2024). Climate change consequences as a topic is associated with lower civility, inclusiveness, and common good reference. Somewhat counter to expectations, political content is positively associated with

civility (cf. Anderson & Huntington, 2017). Follower count shows negative effects on rationality, interactivity, and constructiveness.
The researcher-level random intercept explains an exceptionally high share of variance in common good reference (ICC = 0.613), indicating that certain researchers attract climate-skeptical commentary regardless of platform. Most strikingly, when the parent message is authored by the researcher themselves, rationality, interactivity, civility, and constructiveness are all lower indicating that researcher participation suppresses rather than elevates deliberative quality. Finally, female researchers attract significantly less rational, interactive, and constructive responses.

Discussion and Conclusions
Our research shows that platform choice matters in some regards, but its effects are modest.
Topic, gender, and follower count effects offer little room for strategic adjustment. Most consequentially, direct researcher participation in comment threads appears to attract lower-quality responses, challenging practitioner guidelines that advocate two-way engagement as best practice.
Deliberative reciprocity (the finding that a message' s quality is the strongest predictor of the quality it receives in return) is both the most theoretically significant result and the most actionable one: the quality of a conversation is shaped less by who participates than by how they participate. Theoretically, we hope this paper demonstrates the value of deliberative theory for PES research and invites further debate, including about dimensions we have omitted (notably affective ones) and extensions to other domains and platforms.

References
Anderson, A. A., & Huntington, H. E. (2017). Social media, science, and attack discourse: How twitter discussions of climate change use sarcasm and incivility. Science Communication, 39, 598‑620. https://doi.org/10.1177/1075547017735113
Bates, D., M¨achler, M., Bolker, B., &Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67 (1), 1‑48. https://doi.org/10.18637/jss.v067.i01
C ̂ot´e, I. M., & Darling, E. S. (2018). Scientists on twitter: Preaching to the choir or singing from the rooftops? FACETS. https://doi.org/10.1139/facets-2018-0002
Friess, D., & Eilders, C. (2015). A systematic review of online deliberation research. Policy & Internet, 7 (3), 319‑339. https://doi.org/10.1002/poi3.95
Naab, T. K., Ruess, H.-S., & K¨uchler, C. (2023). The influence of the deliberative quality of user comments on the number and quality of their reply comments. New Media & Society, 146144482311721. https://doi.org/10.1177/14614448231172168
Niemeyer, S., Veri, F., Dryzek, J. S., & B¨achtiger, A. (2024). How deliberation happens: Enabling deliberative reason. American Political Science Review, 118, 345‑362. https://doi.org/ 10.1017/s0003055423000023
Weingart, P., Joubert, M., & Connoway, K. (2021). Public engagement with science−origins, motives and impact in academic literature and science policy. PLOS ONE, 16, e0254201. https://doi.org/10.1371/journal.pone.0254201

Hey ChatGPT, should I quit my job? – How the complexity of conversations with GenAI affects "wise" reasoning and enables human-AI synergy

Christina M. Brunner | LMU Munich; Franziska L. Salzberger | LMU Munich; Katharina G. Kugler | LMU Munich; Felix C. Brodbeck | LMU Munich

Generative Artificial Intelligence (genAI) has not just arrived at most workplaces by now (Bick et al., 2026), but is increasingly prompted for advice on complex personal life situations, such as relationships, mental health, or ethical dilemmas (McCain et al., 2025; UK Dept. for Science & UK Dept. for Culture, 2026). Beyond individual

user queries, online public discourse in social media posts, blogs, or newspapers (co-)created with genAI (Spennemann, 2025; Sun et al., 2025) increasingly shapes the way in which people give and receive life advice (Kumar et al., 2025).
In this work, we explore a "best-case scenario" of human-AI dyads co-creating life advice that is wiser (i.e., more knowledgeable, balanced, context- and value-sensitive; Grossmann et al., 2020; Staudinger et al., 1994) than both an independent human response and a zero-shot AI response. This standard is captured by the concept of synergy: A "gain in performance that is attributable in some way to […] interaction" (Larson, 2010, p. 5), frequently operationalized in human-AI interaction as a joint outcome that supersedes both the independent human and independent AI ability (Vaccaro et al., 2024). Under which conditions can human-AI co-generated life advice be wiser than either agent would have achieved independently?

We propose integrative complexity (IC) – the degree to which different perspectives on a potentially complex matter are distinguished (differentiation) and related to one another (integration; Suedfeld et al., 1992) – in human-AI conversations as a key mechanism that enables human-AI synergy in wise reasoning. Although IC originated as a measure for the complexity of information processing in human individuals or groups (Brodbeck et al., 2021; Coman & Fodor, 2022), it is applicable to any kind of language-based communication (Baker-Brown et al., 1992), such as historical documents, political speeches, performance-related essays, or social media comments (Gruenfeld & Hollingshead, 1993; Jakob et al., 2023; Suedfeld, 2010). Therefore, in the following, we treat and empirically investigate IC as a property of human-AI conversations. In particular, because IC is conceptualized as a context-sensitive state (Suedfeld et al., 1992) that has been successfully manipulated in human interaction (Brodbeck et al., 2021; Moore & Tenbrunsel, 2014), we propose an equivalent procedure to increase the level of IC in human-AI interaction. Thus, we aim to establish IC as a promising leverage point for designing effective human-AI interaction procedures.

We argue that higher levels of IC should foster synergy in human-AI wise reasoning, i.e., enable dyads to give wiser life advice than either agent would have without interaction. First, the differentiation of relevant aspects identified by each agent is a prerequisite for the dyad to benefit from the unique information base and capabilities of each agent ("complementarity", Hemmer et al., 2025). Second, integration should foster the synthesis of differentiated aspects into a coherent, whole, more balanced, and nuanced ("wiser") viewpoint, e.g., by recognizing mutual trade-offs, dependencies, or conceptual relations between different aspects, or by structuring them under an overarching principle or system.
Therefore, in summary, we aim to answer the following research questions: How can IC be promoted in human-AI interaction? And does high IC lead to synergy in human-AI wise reasoning?

Method Procedure. To answer the aforementioned research questions, we conducted an empirical, experimental online study. Participants' main task was to formulate their thoughts on an open-ended, complex life dilemma that was deliberately ill-defined, value-laden, and resistant to straightforward solutions (Staudinger et al., 1994). Such scenarios typically involve fundamental life decisions (e.g., conflicts between personal goals and social obligations, existential crises, or interpersonal tensions) and are framed in ways that require participants to reason beyond factual knowledge. The task was presented in a third-person format (advising someone else). After providing their answer to this scenario, participants answered a questionnaire on further exploratory measures and demographic information.

Design. The study followed a 3 × 2 between-subjects design with varying agents (Human/AI condition, human condition, AI condition) and IC support (control condition vs. IC intervention condition). Participants were randomly assigned to one of the eligible conditions.
In the *IC control conditions*, human participants answered the scenario without further instruction, AI agents answered the scenario based on a simple zero-shot prompt, and humans in the human/AI condition were

provided a chat interface via ResearchChatAI (Becker et al., 2026) where they could freely interact with an AI model (“Mistral Small”, Mistral AI, 2026) to formulate their answer.
In the IC intervention conditions, the interaction process was divided into two sequential phases designed to induce IC. First, participants and AI agents were instructed to independently generate different approaches for dealing with the situation (promoting differentiation). Second, participants and AI agents were instructed to integrate (e.g., weigh or combine) the differentiated aspects. This intervention mirrors previous manipulations of IC in human individuals (Moore & Tenbrunsel, 2014) and groups (Brodbeck et al., 2021).

Measures. We assessed the IC of human-AI interaction, as evidenced in recorded chat logs, using the standardized IC coding manual (Baker-Brown et al., 1992). Possible scores range from 1 (no differentiation) to 7 (full differentiation and integration). As the main dependent variable, we assessed wise reasoning performance in final statements about the scenario based on the Berlin Wisdom Paradigm coding scheme, which specifies objective criteria for “wise” reasoning based on expert ratings (Staudinger et al., 1994). Possible scores range from 5 to 25.

Participants. Based on an a priori power analysis, we aimed for at least 246 responses (41 per condition, which includes 82 single-shot AI outputs) to detect small to medium effects in a 3 × 2 design ($f = .20$, $\alpha = .05$, $1-\beta = .80$). The final sample included 191 human participants (61 % female; mean age 35.48 years; SD = 14.44) and 100 single-shot AI outputs.

Results
Observed levels of IC in human-AI interaction were higher in the intervention condition (M = 4.16, SD = 1.07) than in the control condition (M = 3.55, SD = 1.12), t(55) = 1.98, p = .026, d = 0.57, 95% CI [-0.02; 1.13]. IC was positively related to human-AI wise reasoning performance, r = .32, p = .014, 95% CI [.07, .54]. A mediation analysis showed that human-AI wise reasoning performance was higher in the IC intervention condition (M = 9.20, SD = 3.23) than the control condition (M = 7.61, SD = 2.46), mediated by higher IC of the interaction (indirect effect = .15, SE = .10, 95% CI [.00, .40]). These results indicate that the intervention successfully increased IC in human-AI interaction, which in turn positively affected wise reasoning in human-AI dyads.
To test for evidence of human-AI synergy at high levels of IC, we compared human-AI wise reasoning performance to the baseline performance of independent humans and independent AI agents in a 3 × 2 analysis of variance (ANOVA). Comparisons to the independent AI baseline allowed no inference that human-AI synergy occurred: There was a main effect of the agent condition, $F(2, 285) = 93.64$, $p < .001$, $\eta^2_p = .40$, whereby independent AI performance (M = 12.66, SD = 1.46) superseded human-AI performance (M = 8.38, SD = 2.95) and independent human performance (M = 8.36, SD = 3.03). Nonetheless, comparisons to the independent human baseline support our proposition that human-AI interaction yields benefits at high levels of IC. There was an interaction effect of the agent × IC conditions, $F(2, 285) = 4.59$, $p = .010$, $\eta^2_p = .03$. In the IC control condition, human-AI wise reasoning performance (M = 7.61, SD = 2.46) was descriptively lower than that of independent humans (M = 8.69, SD = 3.80), t(94) = 1.62, p = .055, d = 0.33, 95% CI [-0.08, 0.74], but in the IC intervention condition, human-AI performance (M = 9.20, SD = 3.23) was significantly higher than that of independent humans (M = 8.04, SD = 2.01), t(93) = 2.14, p = .017, d = 0.44, 95% CI [0.03, 0.86]. The degree to which human-AI dyads outperformed individual humans in the IC intervention condition was positively related to the levels of IC of their interaction, r = .39, p = .016, 95% CI [0.08, 0.63].

Discussion
In this study, we applied the concept of IC to explore the conditions under which human-AI interaction can foster “wiser” judgments than either agent could have achieved independently (human-AI synergy). We designed a structured, guided interaction context to induce higher levels of IC in human-AI interaction, and found that higher IC was a driver of more nuanced, context-sensitive judgments in complex life scenarios. Thus, this work contributes to identifying the preconditions for a constructive culture of debate by demonstrating how

IC – a mechanism associated with constructive conflict resolution and cooperation in human dialogue (Kugler & Coleman, 2020; Park & DeShon, 2018) – can be induced and leveraged in human-AI interaction to support the emergence of wiser judgments, advice, or end-products.

We did not find evidence of human-AI synergy at high levels of IC due to an unexpectedly high baseline for wise reasoning by independent AI agents. However, because wise reasoning can only cause meaningful positive effects when grounded and enacted in “real life” contexts (e.g., if it shapes conversations or guides decisions), we interpret the observed wise reasoning scores of independent AI agents, as reflected in a single, task-specific and context-free output, with caution. In contrast, the comparison of human-AI wise reasoning to the baseline by independent humans showed a more interesting and practically relevant pattern: On the one hand, unstructured interaction with AI might lower human wise reasoning performance, but on the other hand, integratively complex interaction with AI can help humans arrive at wiser conclusions than they would on their own. Thus, fostering IC may be a key leverage point for designing constructive human-AI interactions that not only buffer potentially negative effects of interacting with AI but also enable additional benefits to yield wiser end products than humans would have achieved independently.

Future work might explore more specifically how AI can be used to motivate and enable integratively complex and wise online discourse about complex life issues across different online platforms. For example, an open question is how AI agents can be trained or how user interfaces can be designed (e.g., to prompt independent deliberation of both agents and the structured integration of the results) to foster IC without compromising user experience.

In summary, we show that increasing IC in human-AI interactions about complex life questions could be a promising way to reach “wiser” conclusions. Thus, human-AI co-generated online content offering life advice or discussing complex life issues (e.g., blogs, social media posts, comments, or text messages) can be more constructive, nuanced, knowledgeable, and context-sensitive than independently human-generated content if it arises from integratively complex human-AI interaction.

References

Baker-Brown, G., Ballard, E. J., Bluck, S., de Vries, B., Suedfeld, P., & Tetlock, P. E. (1992). The conceptual/integrative complexity scoring manual. In C. P. Smith (Ed.), Motivation and Personality: Handbook of Thematic Content Analysis (401–418). Cambridge University Press.

Becker, M., Jong, D. de, Briker, R., Mennens, K., Heller, J., Mahr, D., & Grewal, D. (2026). Using customized, conversational AI agents in leadership and management research: Benefits, practical illustrations, and best practices. The Leadership Quarterly, 37(3), 101952. https://doi.org/10.1016/j.leaqua.2026.101952

Bick, A., Blandin, A., & Deming, D. J. (2026). The rapid adoption of generative AI. Management Science, Article mnsc.2025.02523. Advance online publication. https://doi.org/10.1287/mnsc.2025.02523

Brodbeck, F. C., Kugler, K. G., Fischer, J. A., Heinze, J., & Fischer, D. (2021). Group-level integrative complexity: Enhancing differentiation and integration in group decision-making. Group Processes & Intergroup Relations, 24(1), 125–144. https://doi.org/10.1177/1368430219892698

Coman, A., & Fodor, O. C. (2022). Antecedents and consequences of group cognitive complexity: a conceptual review. Team Performance Management, 28(3/4), 238–259. https://doi.org/10.1108/TPM-09-2021-0069

Grossmann, I., Weststrate, N. M., Ardelt, M., Brienza, J. P., Dong, M., Ferrari, M., Fournier, M. A., Hu, C. S., Nusbaum, H. C., & Vervaeke, J. (2020). The science of wisdom in a polarized world: Knowns and unknowns. Psychological Inquiry, 31(2), 103–133. https://doi.org/10.1080/1047840X.2020.1750917

Gruenfeld, D. H., & Hollingshead, A. B. (1993). Sociocognition in work groups: The evolution of group integrative complexity and its relation to task performance. Small Group Research, 24(3), 383–405. https://doi.org/10.1177/1046496493243006

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2025). Complementarity in human-AI collaboration: concept, sources, and evidence. European Journal of Information Systems, 34(6), 979–1002. https://doi.org/10.1080/0960085X.2025.2475962

Jakob, J., Dobbrick, T., & Wessler, H. (2023). The integrative complexity of online user comments across different types of democracy and discussion arenas. The *International Journal of Press/Politics*, *28*(3), 580–600. https://doi.org/10.1177/19401612211044018

Kugler, K. G., & Coleman, P. T. (2020). Get complicated: The effects of complexity on conversations over potentially intractable moral conflicts. Negotiation and Conflict Management Research, 13(3), 211–230. https://doi.org/10.1111/ncmr.12192

Kumar, H., Chahal, J., Zhao, Y., Zhang, Z [Zeling], Wei, A. Z., Tay, L., & Anderson, A. (2025). When AI gives advice: Evaluating AI and human responses to online advice-seeking for well-being. https://arxiv.org/pdf/2512.08937 https://doi.org/10.1145/3772318.3791233

Larson, J. R. (2010). In search of synergy in small group performance. Psychology Press.

McCain, M., Linthicum, R., Lubinski, C., Tamkin, A., Huang, S., Stern, M., Handa, K., Durmus, E., Neylon, T., Ritchie, S., Jagadish, K., Maheshwary, P., Heck, S., Sanderford, A., & Ganguli, D. (2025). How people use Claude for support, advice, and companionship. https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship

Mistral AI. (2026). Mistral Small 4 version [Large language model].

Moore, C., & Tenbrunsel, A. E. (2014). “Just think about it”? Cognitive complexity and moral choice. Organizational Behavior and Human Decision Processes, 123(2), 138–149. https://doi.org/10.1016/j.obhdp.2013.10.006

Park, G., & DeShon, R. P. (2018). Effects of group-discussion integrative complexity on intergroup relations in a social dilemma. Organizational Behavior and Human Decision Processes, 146, 62–75. https://doi.org/10.1016/j.obhdp.2018.04.001

Spennemann, D. H. R. (2025). "Delving into" the quantification of AI-generated content on the internet (synthetic data). https://arxiv.org/pdf/2504.08755

Staudinger, U. M., Smith, J., & Baltes, P. B. (1994). Manual for the assessment of wisdom-related knowledge. Max Planck Institute for Human Development and Education. https://dlc.mpg.de/metadata/mpib_escidoc_33500/5/

Suedfeld, P. (2010). The cognitive processing of politics and politicians: Archival studies of conceptual and integrative complexity. Journal of Personality, 78(6), 1669–1702. https://doi.org/10.1111/j.1467-6494.2010.00666.x

Suedfeld, P., Tetlock, P. E., & Streufert, S. (1992). Conceptual/integrative complexity. In C. P. Smith (Ed.), Motivation and Personality: Handbook of Thematic Content Analysis (393–400). Cambridge University Press.

Sun, Z., Zhang, Z [Zongmin], Shen, X., Zhang, Z [Ziyi], Liu, Y., Backes, M., Zhang, Y., & He, X. (2025). Are we in the AI-generated text world already? Quantifying and monitoring AIGT on social media. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 22975–23005. https://doi.org/10.18653/v1/2025.acl-long.1120

UK Department for Science, Innovation and Technology, & UK Department for Culture, Media and Sport. (2026). AI skills for life and work: General public survey findings. https://www.gov.uk/government/publications/ai-skills-for-life-and-work-general-public-survey-findings/ai-skills-for-life-and-work-general-public-survey-findings

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour, 8(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

RECHT & ETHIK Session 1
Verantwortung und Kontrolle in KI-vermittelten Entscheidungen

19.11.2026, 10:15-12:15

Sprache: Deutsch

Programmkomitee: Niina Zuber | bidt; Julian Nida-Rümelin | bidt & LMU München; Dirk Heckmann | bidt & TU München; Felix Freiling | bidt & FAU Erlangen-Nürnberg

Moderation: Felix Freiling | bidt & FAU Erlangen-Nürnberg

Algorithmisches Ermessen: Verantwortung, Transparenz und Legitimität beim Einsatz von KI in der öffentlichen Verwaltung

Laura Crompton | TH Ingolstadt; Josefine Ruppenthal | TH Ingolstadt

1. KI in der öffentlichen Verwaltung: Zwischen Effizienzversprechen und ethischer Verantwortung

Der Einsatz künstlicher Intelligenz und intelligenter Systeme (KI) in der öffentlichen Verwaltung wird gegenwärtig primär als Antwort auf strukturelle Herausforderungen diskutiert. Prognosen für den Fachkräftemangel in der öffentlichen Verwaltung sind in den letzten Jahren stetig gestiegen: von 730.000 fehlenden Stellen im öffentlichen Sektor bis 2030 (Klier & Stern, 2019) auf 840.000 (Fechter, 2023); der Beamtenbund dbb spricht aktuell von bereits 600.000 (tagesschau.de, 2026). Eine konservative lineare Fortschreibung lässt bis 2030 eine Lücke von rund 945.000 Stellen erwarten. Demografischer Wandel, komplexe Einstellungsverfahren und ein veraltetes Arbeitsbild im öffentlichen Sektor verschärfen diese Entwicklung zusätzlich.

In diesem Sinne kann und sollte KI nicht nur als Effizienzmaßnahme, sondern zunehmend als unabdingbarer Lückenfüller für fehlende Ressourcen gesehen werden. In der Praxis manifestiert sich dieses Umdenken bereits in einer Vielzahl von Verwaltungsprozessen, etwa in der KI-gestützten Bewertung von Sozialleistungsansprüchen oder in datenbasierten Analysesystemen in Sicherheitsbehörden (Eubanks, 2019; OECD, 2024). Den meisten dieser Anwendungsfälle ist gemein, dass KI als Entscheidungsassistenzsystem fungiert: Sachbearbeiter:innen werden durch den Output eines KI-Systems in ihrer Entscheidungsfindung unterstützt, ohne diese vollständig aus der Hand zu geben. KI-basierte Entscheidungsassistenzsysteme sind somit eine direkte Antwort auf den Fachkräftemangel in der öffentlichen Verwaltung. Die Frage ist daher längst nicht mehr, ob, sondern wie wir KI in der öffentlichen Verwaltung einsetzen sollten. Damit stellt sich die Frage, wie KI bestehende Entscheidungsstrukturen und -dynamiken in der öffentlichen Verwaltung verändert und welche technikethischen, moralphilosophischen und verwaltungswissenschaftlichen Implikationen sich daraus ergeben.

In diesem Zusammenhang adressiert dieses Paper zwei grundlegende ethische Herausforderungen, die sich aus dem Einsatz von KI-Entscheidungsassistenzsystemen ergeben: Einerseits entstehen Verantwortungslücken bei Verwaltungsentscheidungen und andererseits fehlen Transparenzpflichten gegenüber den Bürger:innen. Entsprechend ist das Paper in zwei aufeinander aufbauende Teile gegliedert. In einem kurzen ersten Teil werden die grundlegenden Entscheidungsstrukturen der öffentlichen Verwaltung erörtert. Darauf aufbauend argumentiert der zweite Teil, inwiefern sich diese Entscheidungsstrukturen durch den Einsatz von KI verschieben und welche gesellschaftlichen Implikationen sich daraus ergeben.

2. Wie die öffentliche Verwaltung Entscheidungen trifft

Verwaltungsentscheidungen folgen einer geteilten institutionellen Logik. Sie sind zugleich regelgebunden und stark formalisiert: Die Legitimität administrativen Handelns beruht in demokratischen Staaten auf der Bindung

an allgemein gültige Regeln und deren konsistente, nachvollziehbare Anwendung. Ergänzt wird diese Struktur durch die Unterscheidung von Verantwortung und Zuständigkeit. In der Verwaltungspraxis wird Verantwortung durch Zuständigkeit ersetzt, um personale Willkür zu reduzieren und Entscheidungen eindeutig institutionell zuordnen zu können (Moore, 1995; Vollmann, 2021). Innerhalb dieses rigiden Rahmens verfügen Verwaltungsakteur:innen dennoch über Ermessensspielräume, ohne die Verwaltung angesichts wechselhafter Lebensrealitäten nicht handlungsfähig wäre (Lipsky, 1980; Bovens & Zouridis, 2002). Entscheidungen im Verwaltungsvollzug entstehen somit in einem Spannungsfeld zwischen Regelbindung und Ermessen.

An dieser Schnittstelle greifen KI-gestützte Entscheidungsassistenzsysteme auf zweierlei Weise in Entscheidungsprozesse ein. Erstens strukturieren sie Entscheidungen vor. Indem KI-Systeme Optionen priorisieren, Fälle einordnen oder Risiken bewerten, ist der Entscheidungsraum bereits definiert, bevor Sachbearbeiter:innen selbst urteilen. Was als Empfehlung erscheint, kann als Rahmensetzung verstanden werden, in der bestimmte Handlungsoptionen sichtbar werden und andere in den Hintergrund treten. Zweitens beeinflusst die algorithmische Ausgabe das menschliche Urteil auch dort, wo formal Ermessen verbleibt. Forschung zu *Automation Bias* zeigt, dass Menschen KI-Outputs auch dann übernehmen, wenn ihr eigenes Urteil abweicht (Araujo et al., 2020; Crompton, 2021; Crompton, 2022; Coeckelbergh, 2025). Beide Mechanismen wirken zusammen. Der vorstrukturierte Entscheidungsraum verengt das Ermessen, der Automation Bias schwächt die Bereitschaft, diesen Rahmen zu hinterfragen. Hieraus ergeben sich zwei eng verwobene normative Herausforderungen.

2.1. Fehlkopplung 404: Wenn Ermessen und Zuständigkeit auseinanderfallen

In der Moralphilosophie ist Verantwortung an *Agency* (dt. Handlungsträgerschaft) geknüpft; ohne *Agency* keine Verantwortung (Nida-Rümelin, 2011; Shoemaker, 2011). Wie bereits erwähnt, arbeitet die Verwaltung mit Zuständigkeit statt mit Verantwortung. Zuständigkeit weist einer Person eine Entscheidungsaufgabe formal zu, unabhängig davon, ob sie im vollen moralphilosophischen Sinne als autonomes Subjekt gehandelt hat. Das macht Entscheidungen zurechenbar und das System handlungsfähig, klammert die Frage nach *Agency* aber bewusst aus (Moore, 1995; Vollmann, 2021). Hier greift nun das Ermessen, durch das *Agency* im Verwaltungsvollzug wirksam wird: es bezeichnet nicht nur die rechtliche Befugnis, innerhalb eines Rahmens zu entscheiden, sondern auch die epistemischen und moralischen Dispositionen, die dafür notwendig sind, wie bspw. die Fähigkeit relevante Evidenz zu gewichten, Werte situativ abzuwägen, und Einzelfälle in ihrer Besonderheit zu erfassen (Lipsky, 1980; Zacka, 2017; Vredenburgh, 2023).
Diese Konstruktion funktioniert, solange die zuständige Person auch de facto die entscheidende Instanz ist. KI-gestützte Entscheidungsassistenzsysteme untergraben jedoch genau diese Voraussetzung. Durch Vorstrukturierung findet der eigentliche Akt des Abwägens bereits in der Systemgestaltung statt, durch Entwickler:innen, die weder verwaltungsrechtlich gebunden noch demokratisch legitimiert sind. Durch *Automation Bias* wird die Bereitschaft der Sachbearbeiterin diesen Rahmen zu hinterfragen, weiter geschwächt. Die *Agency*, aus der sich Verantwortung begründen würde, verlagert sich, doch die Zuständigkeit bleibt.

So entstehen zwei übereinanderliegende Lücken, die das bestehende Verwaltungsmodell nicht schließen kann. Zwischen *Agency* und Verantwortung klafft im klassischen Modell bereits eine bewusste Lücke, die durch Zuständigkeit institutionell geschlossen wird. Doch durch KI entsteht nun eine zweite Lücke, nämlich zwischen Zuständigkeit und tatsächlichem Entscheidungseinfluss. Für diese zweite Lücke hat das System keine Antwort. Wer hat eigentlich entschieden: Beamt:in oder das System? Kann Zuständigkeit Verantwortung ersetzen, wenn die *Agency*, auf der Verantwortung beruht, systematisch woanders liegt? Die KI-Ethik diskutiert diese Problematik häufig als *Responsibility-Gap* (Matthias, 2004); im Verwaltungskontext greift dieser Begriff jedoch zu kurz, da die Konzepte von Zuständigkeit und Ermessen nicht mehr angemessen abgebildet werden. Außerdem handelt es sich hier nicht um eine vollständige Lücke, sondern vielmehr um eine Fehlkopplung zwischen formaler Zuständigkeit und faktischer Einflussnahme. KI-Systeme integrieren sich in bestehende

Verwaltungsstrukturen, ohne deren grundlegende Logik zu verändern; sie treten als Entscheidungshilfen in Erscheinung, während ihre vorstrukturierende Rolle unsichtbar bleibt.
Algorithmische Entscheidungslogiken „verbergen" sich so hinter menschlichen Entscheidungsträger:innen und entziehen sich genau jener Zurechenbarkeit, die der Verwaltungsapparat durch die Trennung von Verantwortung und Zuständigkeit eigentlich herstellen soll.

2.2. Algorithmische Intransparenz und die Erosion administrativer Begründungspflichten
Ein weiterer Aspekt dieser Herausforderung liegt in der Legitimation von Verwaltungshandeln.
Deliberative Demokratietheorien verstehen Legitimität im Kern als Frage der Rechtfertigbarkeit. Entscheidungen müssen gegenüber Betroffenen begründbar sein. Transparenz und Nachvollziehbarkeit sind damit keine optionalen Zusätze, sondern konstitutive Merkmale legitimen Verwaltungshandelns (Moore, 1995; Bovens & Zouridis, 2002). Viele algorithmische Systeme arbeiten jedoch mit Modellen, deren innere Funktionsweise selbst für Fachleute nur eingeschränkt nachvollziehbar ist. Dies wird in der Debatte um Explainable AI (xAI) vielfach thematisiert (Günther & Kasirzadeh, 2022). Wenn solche Systeme Entscheidungen vorstrukturieren, wird fehlende Transparenz zu einem festen Bestandteil administrativer Prozesse, die sich im Nachhinein nur begrenzt rekonstruieren und deren Einflussgrößen sich nicht verlässlich messen lassen. Wie gravierend das ist, lässt sich am Beispiel der öffentlichen Vergabe verdeutlichen. Wenn in der öffentlichen Verwaltung Leistungen oder Güter beschafft werden sollen, muss hierfür eine Leistungsbeschreibung erstellt werden, die genau beschreibt, welche Anforderungen an Bieter:innen gestellt werden.

Diese Leistungsbeschreibungen werden veröffentlicht, sodass jede:r, der die Anforderungen erfüllt, ein Angebot einreichen kann. Die Angebote der Bieter:innen werden dann geprüft und bewertet, bevor der Zuschlag erteilt wird. Diese Prozesse müssen transparent sein, sodass nach einer Entscheidung die Gründe für und gegen den jeweiligen Zuschlag gegeben werden können. Es wird mit Bewertungstabellen gearbeitet und es werden Begründungen für und gegen Bieter:innen erstellt. Wird nun eine KI eingesetzt, die Bieter:innenangebote vorstrukturiert, werden bestimmte Entscheidungsparameter inhärent intransparent und für die Sachbearbeiter:innen nicht nachvollziehbar. Wie bereits in 2.1. argumentiert, ergibt sich durch *Automation Bias* zusätzlich die Tendenz, dass Menschen KI-Outputs unreflektiert übernehmen, was in diesem Fall dazu führt, dass Sachbearbeiter:innen die intransparenten Entscheidungen der KI übernehmen und Verwaltungsentscheidungen damit die Intransparenz von KI forttragen. Dies widerspricht fundamental den Prinzipien des Verwaltungshandelns.

Beide Problemdimensionen greifen ineinander. Die institutionell unsichtbar gemachte Vorstrukturierung durch KI ist zugleich Ursache einer Verantwortungsverschiebung und Hindernis ihrer demokratischen Kontrolle. Sie berührt damit auch die Möglichkeit der Einzelfallgerechtigkeit, die das deutsche Verwaltungsrecht im Kern gewährleisten soll. Obwohl intelligente Entscheidungsassistenzsysteme in der öffentlichen Verwaltung zunehmend Anwendung finden, fehlt bislang eine systematische, empirisch wie ethisch fundierte Untersuchung dieser Veränderungen der öffentlichen Verwaltung. Diese Lücke zu schließen ist nicht nur wissenschaftlich geboten, sondern auch demokratisch dringlich.

3. Gesellschaftliche Implikationen: Ungleichheit, Vertrauen und digitale Souveränität
Bestehende Konzepte der KI-Ethik adressieren Verantwortungsfragen in allgemeiner Form und sind für den Kontext öffentlicher Verwaltung daher nur bedingt tauglich. Denn die öffentliche Verwaltung kann nicht mit anderen Akteuren (wie bspw. Akteuren aus der Wirtschaft) gleichgesetzt werden. Entscheidungen der Verwaltung müssen sich lückenlos auf eine demokratische Legitimationskette zurückführen lassen und sind unmittelbar grundrechtsgebunden. Außerdem sind Bürger:innen auf das Verwaltungssystem (und damit Verwaltungsentscheidungen) angewiesen – es definiert grundlegend ihre Lebensrealität. Sie können nicht einfach auf andere „Verwaltungsanbieter:innen“ ausweichen, sondern sie sind auf Rechtmäßigkeit, Transparenz und Einzelfallgerechtigkeit administrativer Entscheidungen des einen Verwaltungsapparats

angewiesen. Kurzgefasst: In der Privatwirtschaft ist Verantwortung eine Frage von Vertrag, Markt und Haftung. In der Verwaltung ist sie eine Frage der Rechtsstaatlichkeit und demokratischer Legitimation.
In der KI-Ethik gilt Transparenz als zentrale Prämisse für Vertrauenswürdigkeit. Ein System ist vertrauenswürdig, wenn seine Funktionsweise nachvollziehbar ist (HLEG-AI, 2019). Darüber hinaus bedingt Transparenz Autonomie. Wenn unklar bleibt, warum ein System eine bestimmte Empfehlung generiert hat, kann diese weder sinnvoll hinterfragt werden noch eine davon unabhängige Entscheidung getroffen werden. Transparenz ist damit nicht nur eine epistemische, sondern eine konstitutiv autonomiebegründende Anforderung. Ohne Nachvollziehbarkeit keine informierte Handlungsfähigkeit (Mittelstadt et al., 2016). Dieses Grundkonstrukt lässt sich unmittelbar auf die öffentliche Verwaltung übersetzen, und gewinnt dort eine besondere Schärfe, denn die Grundstruktur der öffentlichen Verwaltung ist auf genau diese Nachvollziehbarkeit ausgelegt. Wenn KI-Entscheidungsassistenzsysteme Entscheidungen vorstrukturieren und dabei algorithmische Logiken einbetten, die weder für Sachbearbeiter:innen noch für Bürger:innen nachvollziehbar sind, werden diese Strukturprinzipien nicht formal abgeschafft, sondern still und leise ausgehöhlt. Vertrauen in die Verwaltung setzt voraus, dass Bürger:innen nachvollziehen können, warum Entscheidungen auf eine bestimmte Weise getroffen wurden und welche Gründe wie gewichtet wurden. Ist das nicht möglich, kann das wichtige Implikationen für das Vertrauen in staatliche Institutionen haben.

Ein besonders wichtiger Aspekt dabei sind strukturelle Ungleichheiten und Machtdynamiken der Betroffenheit solcher vorstrukturierten Entscheidungen. Die beiden beschriebenen Problemdimensionen treffen nicht alle Bürger:innen gleich. Wer über hohe Bildung, rechtliche Kenntnisse oder institutionelles Kapital verfügt, kann Verwaltungsentscheidungen eher hinterfragen, rechtlich anfechten und Begründungsdefizite benennen. Wer über diese Ressourcen nicht verfügt (häufig genau jene, die auf staatliche Leistungen am stärksten angewiesen sind), ist algorithmischen (Fehl-)Entscheidungen besonders ausgesetzt (Eubanks, 2019; O'Neil, 2016). Die Intransparenz KI-gestützter Verwaltungsentscheidungen schränkt Autonomie damit auf zwei miteinander verschränkten Ebenen ein: i) unmittelbar, weil Betroffene eine Entscheidung weder verstehen noch sinnvoll anfechten können, und ii) mittelbar, weil das Perpetuieren bestehender Ungleichheiten die Handlungsspielräume derjenigen weiter verengt, die ohnehin schon weniger Möglichkeiten haben, staatliche Entscheidungen anzufechten.

Doch die Frage der Transparenz endet nicht bei der Nachvollziehbarkeit einzelner Verwaltungsentscheidungen. Sie reicht auch tief in die technischen Voraussetzungen. Die Transparenzfrage führt deshalb unmittelbar zur Frage digitaler Souveränität. Wenn KI-Entscheidungsassistenzsysteme, die in deutschen und europäischen Verwaltungen genutzt werden, auf Architekturen und Trainingslogiken beruhen, die außerhalb dieser speziellen demokratischen Einflussbereiche liegen, verlieren Verwaltungen nicht nur die technische, sondern auch die normative Kontrolle über die Systeme, die ihre Entscheidungen vorstrukturieren. Digitale Souveränität ist damit eine konstitutive Bedingung rechtsstaatlicher Handlungsfähigkeit. Wer die Logik eines Systems nicht kennt und nicht kontrolliert, kann auch nicht dafür einstehen.

4. Fazit

Der Einsatz von KI-Entscheidungsassistenzsystemen in der öffentlichen Verwaltung ist keine Zukunftsfrage mehr, sondern ist gegenwärtige Praxis. Umso dringlicher ist eine systematische ethische und rechtliche Auseinandersetzung mit den genannten Herausforderungen. Wir müssen sicherstellen, dass algorithmische Vorstrukturierung nicht zur unhinterfragten Normalität wird, bevor die demokratischen und rechtsstaatlichen Konsequenzen angemessen durchdacht sind. Die in diesem Paper identifizierten Problemdimensionen zeigen: Es geht dabei nicht um Technikskepsis, sondern um die Bedingungen, unter denen KI in der Verwaltung überhaupt legitim eingesetzt werden kann.

Literaturverzeichnis

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence.*AI & Society, 35*(3), 611–623.

Bovens, M., & Zouridis, S. (2002). From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. *Public Administration Review, 62*(2), 174–184.

Coeckelbergh, M. (2025). AI and epistemic agency: How AI influences belief revision and its normative implications. *Social Epistemology*, 1–13.
https://doi.org/10.1080/02691728.2025.2463385

Crompton, L. (2021). The decision-point-dilemma: Yet another problem of responsibility in human-AI interaction. *Journal of Responsible Technology, 8*, 100013.

Crompton, L. (2022). The problem of AI influence. In V. C. Müller (Ed.), *Philosophy and theory of artificial intelligence 2021* (pp. 137–153). Springer International Publishing.

Eubanks, V. (2019). *Automating inequality: How high-tech tools profile, police, and punish the poor*. Picador.

Fechter, M. (2023). *Studie: Im öffentlichen Dienst fehlen bis 2030 140.000 IT-Fachkräfte*. McKinsey & Company. https://www.mckinsey.de/news/presse/2023-01-25-it-talent-im-publicsector

Günther, M. & Kasirzadeh, A. *(2022).* Algorithmic and human decision making: For a double standard of transparency. *AI & Society*, 37(1), 375–381.

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy artificial intelligence*. European Commission. https://digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Klier, J., & Stern, S. (2019). *Bis 2030 fehlen im öffentlichen Sektor in Deutschland 730.000Mitarbeiter*. McKinsey & Company. https://www.mckinsey.de/news/presse/2019-04-02-diebesten-bitte#/

Lipsky, M. (1980). *Street-level bureaucracy: Dilemmas of the individual in public services*. Russell Sage Foundation.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3*(2).*

Moore, M. H. (1995). *Creating public value: Strategic management in government*. Harvard University Press.

Nida-Rümelin, J. (2011). *Verantwortung*. Reclam.

OECD. (2024). *OECD-Bericht zu künstlicher Intelligenz in Deutschland*. OECD Publishing.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group.

Shoemaker, D. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics, 121*(3), 602–632.

tagesschau.de. (2026). *Beamtenbund dbb sieht wachsende Fachkräfelücke im öffentlichen Dienst*. https://www.tagesschau.de/inland/gesellschaft/dbb-fachkraefte-oeffentlicher-dienst-100.html

Vollmann, K. (2021). Verantwortung, Zuständigkeit und Zuständigkeitsverschiebung. In *Compliance in der öffentlichen Verwaltung*. Springer. https://doi.org/10.1007/978-3-658-33879-4_8

Vredenburgh, K. (2023). AI and bureaucratic discretion. *Inquiry, 68*(4), 1091–1120.

Zacka, B. (2017). *When the state meets the street: Public service and moral agency*. Harvard University Press.

Menschliche Aufsicht von AI Agents nach dem AI Act

Chiara Streitbörger | Universität Potsdam; Marietta Ostendorf | Universität Potsdam

AI Agents[1] übernehmen zunehmend (teil-)autonom Aufgaben für Menschen – sei es als persönlicher Shoppingassistent im Alltag eines Verbrauchers,[2] in der Software-Entwicklung,[3] bei der Koordination der Ernte[4] oder zur unternehmensinternen Prozessoptimierung[5]. Während die seit den 1990er Jahren erforschten „Agenten“[6] meist einer regelbasierten Logik folgten und daher nur einen beschränkten Anwendungsbereich hatten, können AI Agents durch Künstliche Intelligenz (KI), insbesondere Large Language Models (LLMs), diese Beschränkungen überwinden. Diese bestehen im Unterschied zu herkömmlicher KI aus drei Komponenten: der Wahrnehmung, dem Denken („Gehirn“[7]) und der Aktion.[8] LLMs übernehmen dazu ausgehend von den Inputs der Nutzer (Prompt und weiterer Kontext) iterativ das *Reasoning* und *Planning*, indem sie Ziele identifizieren, diese in Einzelbestandteile zerlegen, die nächsten erforderlichen Schritte auf Wahrscheinlichkeit beruhend berechnen und Aufträge an einzelne (Sub-)Agenten delegieren.[9] Zur Ausführung können diese z.B. unter Nutzung von APIs und Protokollen dynamisch auf externe Tools zugreifen. Durch ihre sich verstärkende Autonomie versprechen AI Agents Effizienzsteigerungen und Produktivitätsgewinne, da sie kostenintensive menschliche Handlungen ersetzen können.[10] Gleichzeitig sind mit dem Einsatz von AI Agents auch erhebliche Risiken verbunden.

AI Agents werden als KI-Systeme im Rahmen des europäischen AI Acts[11] reguliert.[12] Durch den risikobasierten Ansatz[13], klare Verbote und präventive Pflichten soll die Verordnung vertrauenswürdige KI fördern und negative Auswirkungen durch KI begrenzen. Unterfallen AI Agents der Kategorie der sog. Hochrisiko-KI-Systeme,

---

[1] Eine einheitliche Definition von AI Agents existiert bis dato nicht; für eine Definition von *Digital assistants*: European Law Institute, Guiding Principles and Model Rules on Digital Assistants for Consumer Contracts, 2025, S. 13: „*Digital assistant means an application utilising algorithmic systems that, based on pre-set parameters, user inputs and data obtained from a range of sources, is designed to take actions to accomplish a set of pre-defined objectives for contractual relations*“.

[2] Siehe dazu etwa *Busch*, German Law Journal 2025, 1367 ff.

[3] Grundlegend dazu *Qian et al.,* ChatDev: Communicative Agents for Software Development.

[4] *Sapkota/Roumeliotis/Karkee*, Information Fusion 126 (2026), 103599, S. 17 (siehe dort Abbildung 11 b).

[5] *Hadfield/Koh*, An Economy of AI Agents, S. 9 f.

[6] Siehe dazu etwa *Castelfranchi*, Artificial Intelligence 1998, 157 ff.; *Ferber*, Multi-Agent Systems, 1999.

[7] Siehe dazu auch *Kolt*, 101 Notre Dame Law Review 2026 (im Erscheinen), S. 13.

[8] *Deng* et al., ACM Computer Surveys 2025, Artikel 182, 4 f.

[9] *Deng* et al., ACM Computer Surveys 2025, Artikel 182, 5; *Hacker*/*Holweg*, A pragmatic approach to regulating AI agents, S. 3.

[10] *Kasirzadeh/Gabriel,* Characterizing AI Agents for Alignment and Governance, S. 7; *Chan et al.,* Infrastructure for AI Agents, S. 2.

[11] Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über künstliche Intelligenz).

[12] Siehe dazu auch die FAQ des EU AI Office: „Während KI-Agenten nach dem KI-Gesetz keine eigene Kategorie von KI sind, reichen die Definitionen eines KI-Systems in Artikel 3 Absatz 1 des KI-Gesetzes und eines GPAI-Modells in Artikel 3 Absatz 63 des KI-Gesetzes aus, um KI-Agenten abzudecken.”, https://ai-act-service-desk.ec.europa.eu/de/faq?combine=ai%20agent&faq_category_id=All.

[13] Vgl. Erwägungsgrund 26 AI Act.

unterliegen Anbieter[14] umfangreichen Regularien.[15] Eine dort normierte Pflicht ist die menschliche Aufsicht gem. Art. 14 AI Act, die als zentrale Schutzmaßnahme für Hochrisiko-KI-Systeme gilt.[16] Die zunehmende Autonomie von AI Agents stellt die menschliche Aufsicht vor grundlegende Herausforderungen: Wirksame Aufsicht setzt voraus, dass Menschen AI Agents überwachen und erforderlichenfalls in ihr Agieren eingreifen können. Angesichts der hohen Verarbeitungsgeschwindigkeit, Interaktionen mit verschiedenen AI (Sub-)Agents, ihrer Opazität, der technischen Komplexität der Systeme und der vielfältigen Lösungsmöglichkeiten von Aufgaben wird dies jedoch erheblich erschwert.[17] Es ist daher fraglich, ob die menschliche Aufsicht für AI Agents zweckmäßig ist[18] und ob Menschen diese in der Praxis faktisch ausüben können oder nur formal ‚in' bzw. ‚on the loop' sind. Wird dies verneint, wird menschliche Aufsicht zu einer Scheinverantwortung, die zugleich eine Verantwortungslücke aufwirft. Die Verantwortung wird organisatorisch dem Menschen zugeschrieben, während sie praktisch kaum mehr wahrgenommen wird.

Der Beitrag untersucht daher folgende Fragen: Welche Pflichten haben Anbieter und Betreiber von AI Agents hinsichtlich menschlicher Aufsicht nach dem AI Act? Sind die Anforderungen des Art. 14 AI Act an die menschliche Aufsicht beim Einsatz von zunehmend autonom agierenden AI Agents erfüllbar? Und wenn dies anzunehmen ist: unter welchen technischen bzw. prozessualen Voraussetzungen wird menschliche Aufsicht über AI Agents zu echter Verantwortungsübernahme, und wann bleibt sie bloß eine Fiktion?

I. Konzeptioneller Rahmen und methodische Vorgehensweise

Der Beitrag stellt eine rechtliche Untersuchung auf Grundlage des AI Acts dar. Nach einer kurzen technischen Grundlegung fokussieren sich die Vortragenden auf folgende Aspekte:

1. Rechtliche Einordnung von AI Agents unter dem AI Act: Der Beitrag ordnet zunächst AI Agents als KI-Systeme gem. Art. 3 Nr. 1 AI Act ein und erläutert den risikobasierten Ansatz des AI Acts.
2. AI Agents als Hochrisiko-KI-Systeme: Sodann analysiert der Beitrag die Einstufung von AI Agents als Hochrisiko-KI-Systeme (Art. 6 ff. AI Act) und zeigt die Pflichten der Anbieter und Betreiber anhand von Anwendungsfällen (z.B. Recruiting- oder Kreditvergabe-Agenten) auf.
3. Erfüllbarkeit der Anforderungen an die Aufsicht gem. Art. 14 AI Act: Es werden zentrale Friktionen zwischen Autonomie und Aufsicht herausgearbeitet und technische, organisatorische und prozessuale Bedingungen untersucht, die menschliche Aufsicht über AI Agents zu echter Verantwortungsübernahme machen.

II. Problemfelder

Art. 14 Abs. 1 AI Act: Strukturelle Grenzen der ‚Human in/on the loop'-Konzepte: Art. 14 Abs. 1 AI Act verlangt, dass die Hochrisiko-KI-Systeme so konzipiert und entwickelt werden, dass sie während der Dauer ihrer Verwendung von natürlichen Personen wirksam beaufsichtigt werden können. Die Fähigkeit einer menschlichen Aufsichtsperson, in jeden Entscheidungszyklus des Systems einzugreifen (sog. ‚Human in the loop'), ergänzend dazu den Systembetrieb zu überwachen (‚Human on the loop') oder die Gesamtsteuerung über das System wahrzunehmen (‚Human in command')[19], erscheint schon aufgrund der Konzeption von AI Agents nur schwer umsetzbar. AI Agents basieren häufig auf komplexen Modellen, die Entscheidungen weder

---

[14] Art. 3 Nr. 3 AI Act.

[15] Art. 8 ff. AI Act

[16] Siehe dazu *Dienes*, MMR 2024, 456 (457 ff.).

[17] *Gardhouse*/*Oueslati*/*Kolt*, Regulating AI Agents, S. 41 f.

[18] Es wird teilweise hinterfragt, ob bei AI Agents eine menschliche Aufsicht überhaupt sinnvoll ist, siehe etwa: *Zhu et al.*, AI and Ethics 2026; zurückhaltend *Fleischmann*, RDi 2025, 193 Rn. 27.

[19] HEG-KI, Ethik-Leitlinien für eine vertrauenswürdige KI, 2019, Rn. 65.

linear noch vollständig erklärbar treffen.[20] Sie führen Aufgaben aus, ohne dass jeder Schritt explizit nachvollziehbar ist.

Art. 14 Abs. 2 und Abs. 4 lit. b) AI Act: Risikominimierung und Automation Bias: Art. 14 Abs. 2 AI Act verlangt, dass durch die menschliche Aufsicht Risiken für Gesundheit, Sicherheit oder Grundrechte minimiert werden. Für die Zwecke sollen sie gem. Art. 14 Abs. 4 lit. b) AI Act so zur Verfügung gestellt werden, dass natürliche Personen, denen die menschliche Aufsicht übertragen wurde, angemessen und verhältnismäßig in der Lage sind, sich einer möglichen Neigung zu einem automatischen oder übermäßigen Vertrauen in die von einem Hochrisiko-KI-System hervorgebrachte Ausgabe bewusst zu bleiben. In der Praxis gestaltet sich dies jedoch schwierig, wenn AI Agents eigenständig handeln und Entscheidungen schnell treffen, sodass eine menschliche Kontrolle nicht mehr (hinreichend) möglich ist. Dies gilt insbesondere hinsichtlich der Gefahr von sog. *Automation Bias*: Dies meint die Tendenz von Menschen, automatisierten Vorschlägen mehr Vertrauen zu schenken und widersprüchliche Informationen zu ignorieren, selbst wenn diese korrekt sind.[21] Dadurch können fehlerhafte oder diskriminierende Entscheidungen oftmals unentdeckt bleiben.

Art. 14 Abs. 3 AI Act: Designpflichten und praktische Umsetzbarkeit: Art. 14 Abs. 3 AI Act konkretisiert die Aufsicht dahingehend, dass die Anbieter von AI Agents vor deren Inverkehrbringen oder der Inbetriebnahme geeignete Vorkehrungen treffen müssen, um sicherzustellen, dass eine menschliche Aufsicht gewährleistet ist. Dazu gehören z.B. verständliche Benutzeroberflächen für den Menschen, klare Warnhinweise sowie Mechanismen zur Kontrolle der Ergebnisse der KI. In der Praxis ist dies für AI Agents jedoch nur schwer umsetzbar, da sie viele Einzelentscheidungen treffen, externe Tools nutzen und sich dynamisch an neue Situationen anpassen.[22] Die dadurch entstehenden Risiken werden durch kognitive Verzerrungen wie etwa *Confirmation Bias* und *Anchoring* verstärkt: Menschen neigen dazu, algorithmische Ausgaben als überlegen wahrzunehmen, Entscheidungen unkritisch zu übernehmen und Eingriffe hinauszuschieben, was die Aufsicht weiter schwächt.

III. Thesen

Erste These (Erfordernis echter Aufsicht): Menschliche Aufsicht im Sinne des AI Act darf nicht als bloße „Anwesenheit eines Menschen im Loop“ verstanden werden. Sie setzt vielmehr voraus, dass die Aufsichtspersonen technisch und organisatorisch in der Lage sind, das KI-System zu verstehen,[23] kritisch zu bewerten und wirksam zu begrenzen. Beim Einsatz von AI Agents werden diese Voraussetzungen regelmäßig nicht hinreichend vorliegen.

Zweite These (Faktische Unkontrollierbarkeit): Strukturelle Eigenschaften von AI Agents erschweren die Umsetzbarkeit menschlicher Aufsicht, etwa aufgrund der Opazität der Entscheidungsprozesse von AI Agents, der hohen Verarbeitungsgeschwindigkeit und der Vielzahl an Interaktionen in Multi-Agenten-Systemen. Hinzu kommen agentenspezifische Risiken wie etwa indirekte Prompt Injection Attacks[24], die zu einer Kompromittierung des Systems führen können, ohne dass dies für eine Aufsichtsperson erkennbar ist.

Dritte These (Scheinverantwortung): Die menschliche Aufsicht bei AI Agents droht zu einer Scheinverantwortung zu werden, bei der eine natürliche Person tatsächlich nicht in der Lage ist, das Systemverhalten zu kontrollieren.

---

20 Vgl. Sapkota/Roumeliotis/Karkee, Information Fusion 126 (2026) 103599, S. 9.
21 Art. 14 Abs. 4 lit. b. Siehe zu dem Begriff: Zuiderveen Borgesius, 2018, S. 8.
22 *Sapkota/Roumeliotis/Karkee*, Information Fusion 126 (2026) 103599, S. 4 f.
23 Zu den Anforderungen an das Verständnis: *Dienes*, MMR 2024, 456 (459).
24 Siehe dazu *Greshake* et al., AISec '23: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security 2023, 179 ff.

IV. Literaturverzeichnis

Busch, Christoph	Consumer Law for AI Agents, German Law Journal 2025, 1367-1382.
Castelfranchi, Cristiano	Modelling social action for AI agents, Artificial Intelligence 1998, 157, DOI: 10.1016/S0004-3702(98)00056-3
Chan, Alan et al.	Infrastructure for AI Agents, Artificial Intelligence 2025, https://doi.org/10.48550/arXiv.2501.10114
Deng, Zehan et al.	AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways, ACM Computing Surveys 2025, Artikel 182, 1-36.
Europäische Union	Ethik-Leitlinien für eine vertrauenswürdige KI, 2019, https://data.europa.eu/doi/10.2759/22710.
European Law Institute	Guiding Principles and Model Rules on Digital Assistants for Consumer Contracts, 2025, https://www.europeanlawinstitute.eu/projects-instruments/instruments/eli-guiding-principles-and-model-rules-on-digital-assistants-for-consumer-contracts/
Ferber, Jacques	Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence, 1999
Fleischmann, Christian	KI-Agenten-Systeme: Einordnung in die KI-Verordnung, RDi 2025, 193-199.
Gardhouse, Kathrin/ Oueslati, Amin/ Kolt, Noam	Regulating AI Agents, Computer and Society 2026, https://doi.org/10.48550/arXiv.2603.23471
Greshake, Kai et al.	AISec '23: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security 2023
Hacker, Philipp/ Holweg, Matthias	A pragmatic approach to regulating AI Agents, 2026, https://arxiv.org/abs/2604.22819
Hadfield, Gillian/ Koh, Andrew	An Economy of AI Agents, 2025, https://arxiv.org/abs/2509.01063v1.
Kasirzadeh, Attosa/Gabriel, Iason	Characterizing AI Agents for Alignment and Governance, Computer and Society 2025, https://doi.org/10.48550/arXiv.2504.21848
Kolt, Noam	Governing AI Agents, 101 Notre Dame Law Review (im Erscheinen) 2026
Qian, Chen et al.	ChatDev: Communicative Agents for Software Development, Software Engineering 2023

Riedl, Mark/ Desai, Deven	AI Agents and the Law, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 2025, 2189-2198.
Rothschild, David et al.	The Agentic Economy, 2025, https://arxiv.org/abs/2505.15799v2.
Santoni de Sio, Filippo/ van den Hoven, Jeroen	Meaningful Human Control over Autonomous Systems: A Philosophical Account, Frontiers in Robotics and AI 2018.
Sapkota, Ranjan/ Roumeliotis, Konstantinos/ Karkee, Manoj	AI Agents vs. Agentic AI: A Conceptual taxonomy, applications and challenges, Information Fusion 2026, 103599, 1-30.
Yousefi, Yasaman/ Billi, Marco/ Rotolo, Antonino	Agentic AI: An EU AI Act Paradigm Shift?, 2025, http://dx.doi.org/10.2139/ssrn.5731424.
Zhu, Liming/Lu, Qinghua/ Ding, Ming/Lee, Sung Une/Wang, Cheng	Designing meaningful human oversight in AI, AI & Ethics 2026, https://doi.org/10.1007/s43681-026-01147-7.
Zuiderveen Borgesius, Frederik	Discrimination, artificial intelligence, and algorithmic decision-making, 2018

Künstliche Intelligenz im Online-Handel – Herausforderungen für den Konsumentenschutz

Lena Müller-Naendrup | ÖIAT, Österreich; Karl Gladt | ÖIAT, Österreich

Problemaufriss

Mit der schnellen Verbreitung großer Sprachmodelle (LLMs) und der darauf aufbauenden Entwicklung teilautomatisierter KI-Agenten verändert sich der Online-Handel grundlegend. KI-Chatbots wie ChatGPT, Gemini oder Perplexity treten zunehmend an die Stelle klassischer Suchmaschinen und Vergleichsportale, indem sie auf natürlichsprachliche Anfragen mit aggregierten, vermeintlich kuratierten Antworten reagieren (Grollmann, 2025; McRoskey et al., 2026; Schumacher et al., 2026). Anders als Suchmaschinen oder Online-Plattformen, die auf Basis von Suchbegriffen eine Liste verlinkter Ergebnisse ausgeben, ermöglichen KI-Chatbots eine dialogbasierte Interaktion: Nutzer:innen können direkte Fragen stellen, ihre Anforderungen iterativ präzisieren und erhalten extrahierte Antworten aus unterschiedlichen Online-Shops und Preisvergleichsportalen, ohne das Dialogfenster verlassen zu müssen. Eine seit Herbst 2025 verfügbare Funktion in ChatGPT erlaubt darüber hinaus Nutzer:innen in den USA bereits den unmittelbaren Kaufabschluss im Dialogfenster, ohne externe Online-Shops oder Zahlungsdienste anzusteuern (OpenAI, 2025). Auszugehen ist davon, dass künftig ähnliche Entwicklungen auch Europa erreichen werden.

Als nächster Innovationsschritt im Bereich der Künstlichen Intelligenz gilt das Aufkommen sogenannter *AI Agents*. Darunter werden Softwareprogramme verstanden, die in der Lage sind, eigenständig Aktionen im digitalen Raum auszuführen, mit externen Diensten und Programmierschnittstellen (APIs) zu interagieren und mehrstufige Aufgaben zu planen. Kommerzielle agentische KI Systeme wie Open AI Operator oder Perplexity Comet sind seit Mitte 2025 innerhalb der EU verfügbar und versprechen, Produktinformationen autonom einzuholen, gegen artikulierte Präferenzen abzugleichen und Bestellvorgänge im Auftrag der Konsument:innen vorzubereiten oder direkt abzuschließen (Allouah et al., 2025). Zukünftig könnten AI Agents das manuelle Browsen und die Interaktion mit Online-Plattformen durch automatisierte Selektions- und Bestellvorgänge

weitgehend ersetzen: Es genügt dann, den Bedarf im Dialogfeld eines Chatbots anzugeben, und das Agentensystem navigiert selbstständig durch Online-Shops und Produktkataloge, holt Angebote ein, bewertet sie anhand hinterlegter Präferenzen und wickelt anschließend den Kauf ab.

Dadurch verändert sich die Rolle der Konsument:innen. Während sie bei der klassischen Produktsuche über Suchmaschinen oder Vergleichsplattformen die zentralen Entscheidungsschritte selbst vollziehen, wird ihre Aufgabe bei agentenbasiertem Einkaufen auf die Auftragsformulierung reduziert; die eigentliche Suche, Auswahl und zum Teil bereits das Kaufen selbst werden an KI-Agenten ausgelagert. Die Auftragsformulierung im Dialog tritt an die Stelle der bisherigen Auswahl aus Ergebnislisten, und die Versprechen der Anbieter (Zeitersparnis, individuelle Kuratierung, bessere Entscheidungen) stehen in einem noch ungeklärten Verhältnis zu dem, was die Systeme tatsächlich leisten.

Erste empirische Befunde deuten darauf hin, dass diese Verlagerung mit normativen Implikationen einhergeht. Allouah et al. (2025) zeigen für agentisches E-Commerce, dass Kaufentscheidungen von KI-Agenten teils erheblich von den Zielangaben der Auftraggeber:innen abweichen: Modellgewichtungen wie Position in der Trefferleiste oder Bewertungssignale überlagern artikulierte Präferenzen; mangelnde Rationalität der Modelle führt zu suboptimalen, mitunter teureren Kaufentscheidungen. Hinzu kommt, dass große Sprachmodelle strukturell dazu neigen, Nutzer:innen zu gefallen. Sie bestätigen Annahmen, spiegeln Präferenzen zurück und widersprechen selten. Zu erwarten ist, dass diese als Sycophanz bezeichnete Tendenz auch impulsive Kaufentscheidungen eher bekräftigt als hinterfragt (Cheng et al., 2026). Kommerzielle Geschäftsmodelle wie Affiliate-Provisionen oder Sponsored Placements können Empfehlungen unterhalb der Wahrnehmungsschwelle zusätzlich verzerren (Maes, 2026; Mukherjee & Chang, 2025). Konsument:innen erhalten also an verschiedenen Stellen unbemerkt von ihrer Auftragsvorgabe abweichende Ergebnisse, ohne dass diese Einflüsse für sie sichtbar oder überprüfbar wären.

Daraus erwächst ein Bündel praxisrelevanter Risiken: finanzielle Nachteile durch verzerrte Kaufempfehlungen, Manipulationsrisiken durch für Konsument:innen nicht sichtbare kommerzielle Verflechtungen, Einschränkungen der eigenen Entscheidungs- und Vergleichskompetenz sowie unklare Verantwortlichkeiten bei fehlerhafter Auftragsausführung, da sich die Entscheidungskette auf mehrere Akteure (Konsument:in, Modellanbieter, Plattform, Händler:in) ohne klare Zurechnungsachse verteilt. Empirische Befragungen zeigen gleichwohl ein hohes und schnell wachsendes Vertrauen in KI-gestützte Empfehlungen, das mit der Nähe der KI zum Händler sogar zunimmt (Cheris et al., 2025; McRoskey et al., 2026), also gerade dort am höchsten ist, wo strukturelle Interessenskonflikte am wahrscheinlichsten sind. Konsument:innen delegieren damit Entscheidungen an Systeme, deren Verhalten sie weder vollständig verstehen noch zuverlässig überprüfen können – und deren Anbieter zugleich eigene kommerzielle Beziehungen zu Händler:innen, Plattformen und Werbenden unterhalten. Es steht damit zur Debatte, ob KI-Agenten tatsächlich im Sinne der Auftragsperson handeln und sichergestellt werden kann, dass diese nicht im Interesse Dritter agieren.
Diese Entwicklungen sind keine abstrakten Zukunftsvisionen: Internationale Anbieter wie Perplexity und europäische Online-Händler wie der österreichische Lieferdienst Gurkerl stellen bereits KI-gestützte Shopping-Umgebungen mit integrierten Assistenzfunktionen für europäische Nutzer:innen bereit. Sie machen die im internationalen Diskurs verhandelten Risiken für europäische Konsument:innen konkret und bieten eine alltagsnahe Untersuchungsumgebung, an der sich offene Fragen des Konsumentenschutzes exemplarisch nachvollziehen lassen.

Für den Konsumentenschutz entsteht daraus eine Reihe neuer Aufgabenfelder: In der Aufklärung über tatsächliche Funktionsweisen und Grenzen von KI-Empfehlungen, im Beratungs- und Beschwerdemanagement bei agentenvermittelten Käufen sowie in der Bearbeitung von Informationsasymmetrien zwischen Konsument:innen, KI-Anbietern, Plattformen und Händler:innen. Parallel werden in der aktuellen Diskussion mehrere regulatorische Rahmen als Anknüpfungspunkte gehandelt: darunter der EU AI Act, der Digital

Services Act, das Gesetz gegen unlauteren Wettbewerb (UWG), der Digital Market Act und Verbraucher-Richtlinien - deren Reichweite gegenüber agentischer Kaufabwicklung jedoch erst noch zu klären ist.
Im Horizont eines Digitalen Humanismus geht es somit nicht allein um die abstrakte Kompatibilität von KI-Shopping-Systemen mit demokratischen Grundwerten, sondern um die praktische Frage, ob diese Systeme die Autonomie und die Schutzposition von Konsument:innen aktiv stärken oder schleichend erodieren - und wie der Konsumentenschutz darauf antworten kann.

Forschungsfragen
Aus diesem Problemaufriss leiten sich zentrale Fragestellungen, die im Rahmen des Beitrags beantwortet werden:

1. Welche GenAI-Anwendungen sind im Online-Handel aktuell relevant, und welche Entwicklungen sind ablesbar?
2. Wie verändern sich Such-, Auswahl- und Kaufprozesse für Konsument:innen durch den Einsatz von KI-Chatbots und KI-Shopping-Agenten?
3. Welche Risiken für Konsument:innen lassen sich aus der Untersuchung ableiten?
4. Inwieweit erfassen bestehende rechtliche Rahmenbedingungen agentenvermittelte Kaufabwicklungen und wo bestehen Schutzlücken?

Methodische Vorgehensweise
Der Beitrag berichtet aus einer laufenden, vom österreichischen Sozialministerium geförderten Studie und stütz sich auf einen multimethodischen Zugang:

- Eine **systematische Literaturanalyse** umfasst peer-reviewte wissenschaftliche Beiträge, Reports international tätiger Beratungs- und Marktforschungsunternehmen, Stellungnahmen europäischer und nationaler Verbraucherschutzorganisationen, sowie europäische Konsultationsdokumente, insbesondere zum Digital Fairness Act.
- **Exemplarische Praxistests** untersuchen ausgewählte KI-Anwendungen anhand standardisierter Kaufszenarien. Im Fokus stehen für europäische Konsument:innen unmittelbar verfügbare Systeme, insbesondere die Shopping-Assistenten der Online-Anbieter Gurkerl und Alfies, sowie generische Chatbots (ChatGPT) und Agentensysteme (z.B. Perplexity Comet). Tests werden über Screenshots und Screen Recordings dokumentiert und in nachvollziehbarer Form aufbereitet.
- Eine **rechtliche Analyse** kartiert die als Anknüpfungspunkte gehandelten Regelwerke, darunter der EU AI Act, der Digital Services Act, das Gesetz gegen unlauteren Wettbewerb (UWG), der Digital Markets Act, sowie Verbraucher-Richtlinien und identifizieren offene regulatorische Fragestellungen, die sich aus den Befunden ergeben.

Beitrag zur interdisziplinären Diskussion
Der Beitrag verbindet anwendungsbezogene Marktbeobachtung, empirische Praxistests, eine kartierende rechtliche Analyse und konsumentenpolitische Reflexion und ordnet sich so in die Session *Recht und Ethik* ein. Er nutzt das Feld des KI-gestützten Online-Handels, um die in der Sessionbeschreibung aufgerufenen Fragen nach *Verantwortung, Vertrauen und Integrität* in digitale Ordnungen exemplarisch zu prüfen: Berührt werden dabei Fragen der Zurechenbarkeit über mehrstufige Akteursketten (Konsument:innen, Modellanbieter, Händler:in), der Anwendbarkeit und etwaiger Lücken bestehender Regelwerke sowie der normativen Bewertung von Sycophanz und unsichtbaren Empfehlungsverzerrungen. Zugleich knüpft der Beitrag an Fragestellungen der Sessions *Wirtschaft und Arbeit* (Verschiebung von Markt- und Marketingstrukturen) und *Kommunikation und Gesellschaft* (Informationsasymmetrien gegenüber dominanten Modellanbietern, Kooperationen mit großen Handelsketten) an. Damit nutzt der Vortrag einen Anwendungsfall, der hunderte Millionen europäischer Konsument:innen täglich betrifft, um die Tragfähigkeit etablierter Schutzstrukturen unter Bedingungen teilautonomer KI-Systeme empirisch und normativ zu prüfen.

Literaturverzeichnis
Allouah, A., Besbes, O., Figueroa, J. D., Kanoria, Y., & Kumar, A. (2025). What Is Your AI Agent Buying? Evaluation, Implications and Emerging Questions for Agentic E-Commerce (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2508.02630
Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391(6792), eaec8352. https://doi.org/10.1126/science.aec8352
Cheris, A., Kamel, M.-A., Vu, M., Koszyk, S., Burns, M., & Lewis, R. (2025, November 13). Agentic AI in Retail: How Autonomous Shopping Is Redefining the Customer Journey. Bain. https://www.bain.com/insights/agentic-ai-in-retail-how-autonomous-shopping-redefining-customer-journey/
Grollmann, D. (2025, Oktober 15). *Agentic Commerce: Recherche ja, Kaufabschluss nein*. iBusiness.de - Zukunftsforschung für interaktives Business. https://www.ibusiness.de/aktuell/db/511636grollmann.html
Maes, P. (2026, Januar 25). A Framework for Studying AI Agent Behavior: Evidence from Consumer Choice Experiments. MIT Media Lab. https://www.media.mit.edu/publications/abxlab/
McRoskey, G., Taylor, L., Eppler, B., Barrios, G., Wiener, L., & Lellouche Tordjman, K. (2026). CUSTOMER INSIGHTS Consumers Trust AI to Buy Better. Brands Need to Move Quickly. Boston Consulting Group. https://www.bcg.com/publications/2026/consumers-trust-ai-to-buy-better-brands-must-adapt
Mukherjee, A., & Chang, H. H. (2025). Agentic AI: Autonomy, Accountability, and the Algorithmic Society (arXiv:2502.00289). arXiv. https://doi.org/10.48550/arXiv.2502.00289
OpenAI. (2025, November 25). In ChatGPT einkaufen: Instant Checkout und das Agentic Commerce Protocol. https://openai.com/de-DE/index/buy-it-in-chatgpt/
Schumacher, K., Keutel, M., Kluge, P., Wendler, & Giebel, K. (2026). *Europe's agentic commerce moment: Decision influence is here; execution is coming. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/europes-agentic-commerce-moment-decision-influence-is-here-execution-is-coming*

Das Problem der Verantwortung in Zeiten KI-basierter Herrschaft

Juan Camilo Castillo Rivadeneira | InfAI; Vanita Römer | InfAI

I. Jenseits der positiven Effekte in Bezug auf die versprochene Effizienz- und Qualitätssteigerung gehen mit KI-Systeme eine neue Logik und neue soziopolitische Herausforderungen einher, die sowohl den gesellschaftlichen Zusammenhalt als auch die demokratische Ordnung ins Wackeln bringt. Jenseits seiner theoretischen Dimension befasst sich der Begriff der Verantwortung mit der Relation zwischen Handlungen, ihren (negativen) Folgen und den korrektiven Handlungen (Bayertz, 1995). Vermutlich ihre Fähigkeit diese drei Szenarien zu verbinden gerechtfertigt, dass der Begriff der Verantwortung an Relevanz im Rahmen gesellschaftlichen Transformationsprozessen gewinnt, in den die bereits etablierten Strukturen und Interaktionslogiken sich im Umbruch befinden.
Die Entstehung und die rasante Massifizierung von KI-Systeme –darunter sind die Large Language Models (LLM) die meistbekannten– haben also alle menschliche Lebensbereichen in einer vielfältigen und tiefgreifenden Form transformiert (Bedford-Strohm et al., 2019). Daraus lässt sich begreifen, weshalb der Begriff der Verantwortung dann eine Schlüsselrolle sowohl für die politische Debatte als auch für die wissenschaftliche Forschung zur Erklärung der Relationen und Rollen in der Gesellschaft sowie als Handlungsorientierung für die Mitglieder der politischen Gemeinschaft übernimmt.

In diesem Kontext scheint gleichzeitig allgemein angenommen zu sein, dass zum einen die klassischen Konzeptionen der Verantwortung –Beispiele dafür sind etwa Hans Jonas' (2009) Untersuchung der Verantwortung in Bezug auf die technologische Entwicklung der Atombomben, Valentin Becks (2016) Untersuchung des Verhältnisses zwischen Verantwortung und Weltarmut oder Alexandria Krugs (2020) Auseinandersetzung mit dem Problem der Verantwortung im Kontext der Klimakrise– unzureichend auf die Herausforderungen der KI-

Gesellschaft Antwort geben können. Zum anderen, der Begriff der Verantwortung bleibt wiederum im Mittelpunkt sowohl des politischen Appells für einen *verantwortungsvolle* Umgang (Algorithm Watch / Ch, Juni.2024; Maron, 2024) mit KI-Systeme sowie für wissenschaftliche Analysen über die Rolle und Entscheidungshoheit jeden gesellschaftlichen Akteures im Rahmen der gesellschaftlichen Transformation.

Der aktuelle Beitrag basiert also auf der Prämisse der Zentralität des Verantwortungsbegriffes in der Diskussion um den Umgang mit KI-Big-Tech und die Verteidigung der demokratischen Grundordnung. Darüber hinaus wird hier verteidigt, dass die aktuell laufende gesellschaftliche Transformation die Entstehung eines neuen Gesellschaftsvertrages erforderlich macht, der einen neuen Rahmen für die Interaktion unter demokratischen Prinzipien festlegt. Demnach wird in diesem Beitrag verteidigt, dass es legitim ist, unterschiedlichen Handlungstypen basiert auf differenzierten Verantwortungszuschreibungen für die verschiedenen Akteuren zu erwarten. Die folgende Frage erweist sich als diejenige, die am besten einen Rahmen der Analyse für diesen Beitrag schafft: Welche Verantwortungszuschreibungen sind legitimerweise für gesellschaftliche Akteure im Rahmen der KI-Massifizierung zu erwarten?

II. Dass eine solche Frage zum Programm wissenschaftlichen Untersuchungen zum Thema Verantwortung gehört, lässt sich anhand des Palantir-Beispiels näher begreifen: Das US-Techkonzern Palantir hat neulich auf seinen offizielle X-Account das 22-Punkte Manifest (Palantir, 2026) veröffentlicht. Dabei handelt es sich um eine Zusammenstellung der Thesen, die bereits im Palantir-CEOs [Alexander Karp] Buch *The Technological Republic* (Karp & Zamiska, 2025) zu finden waren. Aus hauptsächlich zwei Gründen hat das Manifest für höhe Resonanz in den Medien gesorgt: Zum einen, es wurde darauf hingewiesen, dass Palantir sich durch das Manifest als aktiver „Akteur an der Schnittstelle zwischen Technologie und Sicherheitspolitik“ (Ehl, 2026) positioniert und, zum anderen, dass es sich dabei um die Reproduktion von Faschistischem Gedankengut des US-Unternehmens handele (Stöcker, 2026). Gleichzeitig haben andere –vielleicht noch wichtigere– Aspekte dennoch wenig bis kaum Aufmerksamkeit erhalten. Hier möchten wir uns auf zwei dieser Aspekte beziehen: einerseits, es verdeutlicht eine klare Selbstkonzeption bzw. ein klares Selbstbewusstsein eines Unternehmens bezüglich seiner strukturellen Rolle und gesellschaftlichen Macht. Andererseits, es betont die Herausforderung, die gewisse legitime Strukturen für die gesellschaftliche Stabilität und die politische Grundordnung verkörpern (Ulbrich & Frey, 2024).

Aus dem Fall Palantirs wäre es möglich viele Fragen herzuleiten, wie zum Beispiel eine Diskussion zu sozioontologischen Aspekten der Unternehmen (Neuhäuser, 2011), die es ermöglicht diese als moralische Akteure wahrzunehmen; die Frage nach der gesellschaftlichen Rolle eines Unternehmens im Rahmen gesellschaftlicher Transformation (Heidbrink, 2012; Kreide, 2013; Polanyi, 2002) bis hin in die Frage nach der Legitimität eines Unternehmens trotz nicht Beibehaltung von minimalen ethischen Standards der demokratischen Ordnung (Niesen, 2013) usw. Anderes gesagt, eine ganze Herausforderungsbreite werden durch den Fall Palantirs geäußert, vor allem, weil die Macht für eine Realitätsbildung nicht nur hypothetisch in den Händen des Unternehmens liegen. Vielmehr Palantir –als paradigmatisches Beispiel dieses Phänomens– verfügt über reale Möglichkeiten dazu, die Realität nach seinem politischen und ethischen Standard zu modellieren und es ist dessen bewusst.

Alle gesellschaftlichen Akteure zu untersuchen, würde den Rahmen dieses Beitrages springen, deswegen werden wir unsere Untersuchung auf zwei Akteure einschränken, nämlich, Staat und Unternehmen. In diesem Vortrag möchten wir uns demnach darauf fokussieren, welche Art von Verantwortungszuschreibung auf Staats- und Regierungsebene möglich ist, um ethische demokratische Prinzipien in den heutigen Zeiten legitimerweise weiterhin verteidigen zu können. Dafür soll nicht nur die neue existierende Machtkonstellation, die durch die rasante Massifizierung von KI-Systeme entstanden ist, und die Rolle von Big-Tech-Konzernen, die staatliche einst demokratisch kontrollierte Bereiche für sich sichern, im Betracht gezogen werden, sondern auch soll eine

Auseinandersetzung mit den möglichen Verantwortungszuschreibung für Unternehmen stattfinden, die den Handlungsspielraum von Unternehmen für die Beibehaltung der demokratischen Ordnung thematisiert. Dieser Beitrag wird auf einer philosophischen Auseinandersetzung mit dem Begriff der Verantwortung und mit den Bedingungen der Möglichkeit für Verantwortungszuschreibungen stehen sowie eine Problematisierung der inhärenten Macht und der intrinsischen Moralität von Big-Tech-Unternehmen. Es wird außerdem die Rolle des Staates als politischer Verantwortungsträger schlechthin analysiert und für den Ansatz argumentiert, dass durch die Überwindung des Palantir-Problems den demokratischen Zusammenhalt gewährleistet werden kann. Das Novum dieses Beitrages besteht darin, aus der philosophischen Perspektive das Verhältnis Macht-Moral-demokratischen Zusammenhang anhand der Verantwortung von Staat und Unternehmen im Kontext der KI konditionierten Gesellschaft zu theorisieren. Gleichzeitig strebt er an, einen Beitrag zu interdisziplinären Diskussionen zu leisten. Der Vortrag findet in deutscher Sprache statt.

References

Algorithm Watch / Ch. (Juni.2024). *Künstliche Intelligenz mit Verantwortung… ohne Diskriminierung: Appell an den Bundesrat* [Press release]. https://algorithmwatch.ch/de/wp-content/uploads/2024/06/Appell_KI-ohne-Diskriminierung.pdf

Bayertz, K. (1995). Eine kurze Geschichte der Herkunft der Verantwortung. In K. Bayertz (Ed.), *Verantwortung: Prinzip oder Problem?* (pp. 3–71). Wissenschaftliche Buchgesellschaft.

Beck, V. (2016). *Eine Theorie der globalen Verantwortung: Was wir Menschen in extremer Armut schulden* (Erste Auflage, Originalausgabe). *Suhrkamp-Taschenbuch Wissenschaft*. Suhrkamp.

Bedford-Strohm, J., Höhne, F., & Zeyher-Quattlender, J. (2019). Digitale Öffentlichkeit zwischen Euphorie und Frustration: Hinführung zur ethischen und politischen Reflexion einer Generationenaufgabe. In J. Bedford-Strohm, F. Höhne, & J. Zeyher-Quattlender (Eds.), *Kommunikations- und Medienethik: Vol. 10. Digitaler Strukturwandel der Öffentlichkeit: Interdisziplinäre Perspektiven auf politische Partizipation im Wandel* (pp. 9–22). Nomos.

Ehl, D. (2026, April 21). *Palantir-Manifest: Kommt nun der "Techno-Faschismus"?* https://www.dw.com/de/palantir-demokratie-militaer-usa-konzern-peter-thiel-ueberwachungssoftware-kuenstliche-intelligenz/a-76884862

Heidbrink, L. (2012). Unternehmen als politische Akteure. Eine Ortsbestimmung zwischen Ordnungsverantwortung und Systemverantwortung / Companies as political actors. Locating regulatory responsibility and systemsresponsibility. *ORDO*, *63*(1), 203–232. https://doi.org/10.1515/ordo-2012-0116

Jonas, H. (2009). *Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation*. Suhrkamp Verlag.

Karp, A., & Zamiska, N. (2025). *The Technological Republic: Über die Macht des Silicon Valley und die Zukunft des Westens*. Finanz Buch Verlag.

Kreide, R. (2013). Die Verantwortung für ein T-Shirt. Individuelle und kollektive Verpflichtungen bei Iris Marion Young. In P. Niesen (Ed.), *Schriftenreihe der Sektion Politische Theorien und Ideengeschichte in der Deutschen Vereinigung für Politische Wissenschaft: Vol. 26. Zwischen Demokratie und globaler Verantwortung: Iris Marion Youngs Theorie politischer Normativität* (1. Aufl., pp. 59–77). Nomos. https://doi.org/10.5771/9783845242958-59

Krug, A. (2020). Das klimaethische Selbst als Subjektivierungsform: Eine idealtypische Konstruktion. *Journal Für Psychologie*, *28*(2), 171–190.

Maron, H.-J. (2024, June 27). *Appell: Bundesrat soll Diskriminierung durch KI verhindern.* Inside IT. https://www.inside-it.ch/appell-bundesrat-soll-diskriminierung-durch-ki-verhindern-20240627

Neuhäuser, C. (2011). *Unternehmen als moralische Akteure*. Suhrkamp Verlag.

Niesen, P. (Ed.). (2013). *Schriftenreihe der Sektion Politische Theorien und Ideengeschichte in der Deutschen Vereinigung für Politische Wissenschaft: Vol. 26*. *Zwischen Demokratie und globaler Verantwortung: Iris Marion Youngs Theorie politischer Normativität* (1. Aufl.). Nomos. https://doi.org/10.5771/9783845242958

Palantir. (2026, April 18). *The Technological Republic, in brief.: Because we get asked a lot [*Post auf Plattform X]. https://x.com/PalantirTech/status/2045574398573453312?s=20
Polanyi, K. (2002). *Chronik der großen Transformation: Wirtschaftliche Transformation, Gegenbewegungen und der Kampf um die Demokratie*. Metropolis-Verlag.
Stöcker, C. (2026, April 26). *So kann die EU niemals wirklich souverän sein* [Meinung - Technologische Unabhängigkeit]. Spiegel Wissenschaft. https://www.spiegel.de/wissenschaft/mensch/manifest-von-palantir-ceo-alex-karp-so-kann-die-eu-niemals-wirklich-souveraen-sein-a-94de7926-bf0a-490b-b4fe-ca97c48c8a50
Ulbrich, C. R., & Frey, B. S. (2024). *Automated Democracy: Die Neuverteilung von Macht und Einfluss im digitalen Staat*. Verlag Herder.

WIRTSCHAFT & ARBEIT Session 2
Governance digitaler Transformation als Mehrebenen-Phänomen: zwischen Organisationen, Institutionen und Wertschöpfungssystemen

19.11.2026, 13:15-15:15

Sprache: Deutsch

Programmkomitee: Angela Graf | bidt; Sabine Pfeiffer | bidt & FAU Erlangen-Nürnberg; Thomas Hess | bidt & LMU München; Alexander Pretschner | bidt & TU München

Moderation: Thomas Hess | bidt & LMU München

Transformation durch Interfaces: Digitale Wertschöpfungsökosysteme und die Verarbeitung pluraler gesellschaftlicher Anforderungen

Benjamin Doubali | Johannes Gutenberg Universität Mainz

Von digitaler Souveränität über Sicherheit und Resilienz inmitten multipler geopolitischer Krisen bis hin zur Klimaneutralität: Industrieorganisationen stehen vor zahlreichen gesellschaftlichen Anforderungen, auf die sie reagieren müssen. Aus soziologischer Sicht lässt sich dieser Verarbeitungsmodus so beschreiben: Externe Erwartungen werden in das Handeln der Organisationen *übersetzt.*
Mit meinem Beitrag untersuche ich, wie Industrieorganisationen gesellschaftliche Anforderungen über digitale Plattformen und Dateninfrastrukturen verarbeiten. Im Mittelpunkt steht die soziologische Frage, wie Organisationen diese heterogenen Erwartungen durch Übersetzungsprozesse und Governance-Mechanismen in technologische Lösungen überführen. Empirisch analysiere ich zwei Fälle industrieller Plattformisierung. Beide Fälle zeigen, wie Interfaces (Schnittstellen) als Medien der Governance wirken und wie sie Übersetzungsprozesse strukturieren, aber auch neue Abhängigkeiten oder Komplexitäten erzeugen. Der Beitrag möchte sozialwissenschaftliche Plattformforschung mit organisationssoziologischen Konzepten verbinden und ein interdisziplinär anschlussfähiges Diskussionsangebot machen, indem er die Ambivalenz digitaler Transformation betont.

Um mit dem Problem vielfältiger und häufig widersprüchlicher Anforderungen an das Handeln von Organisationen umzugehen, verfolgen Industrieorganisationen Lösungsstrategien, die mit technologischer Innovation verbunden sind. So investieren sie etwa in Dateninfrastrukturen und Datenplattformen: Plattformbasierte Zusammenarbeit wird dabei nicht nur als Möglichkeit gesehen, drängende Probleme zu bearbeiten, sondern auch neue Formen industrieller Wertschöpfung zu erschließen (Dolata & Radig, 2025). Mit meinem Beitrag frage ich aus soziologischer Perspektive, wie Organisationen diese Übersetzungsprozesse gestalten und welche Rolle Plattformen und digitale Wertschöpfungsökosysteme dabei für ihre Wandlungsfähigkeit spielen. Dabei ziehe ich die sozialwisschenschaftliche Plattformforschung heran, um mich sowohl mit der Governance neuer digital Wertschöpfungssysteme und Datenräumen, als auch mit „klassischen“ Industrieplattformen zu beschäftigen. Dazu gehören IIoT-Plattformen (Industrial Internet-of-Things) als auch Cloud Services (Dolata, 2024). Die Untersuchung von Plattformen hat sich in den letzten Jahren zu einem Schwerpunkt (soziologischer) Medien- und Digitalisierungsforschung entwickelt (Kenney & Zysman, 2016; Nieborg et al., 2022; van Dijck et al., 2018), da sie als sozio-technische Infrastrukturen nicht nur Interaktionen und Transaktionen reorganisieren, sondern auch neue Formen der sozialen Koordination (Dolata & Schrape, 2025; Gawer, 2022; Kirchner & Beyer, 2016) und Kontrolle hervorbringen (Schaupp, 2021; Staab, 2019; Zuboff, 2018).

Während sich die Plattformforschung bisher vorrangig mit Social Media oder anderen konsumorientierten Internetplattformen beschäftigt (Alaimo & Kallinikos, 2017; Helmond et al., 2019), werden Plattformen, die in spezifische Wissenskulturen und Handlungsfelder eingebettet sind (wie in der Medizin, Wissenschaft oder

Industrie) bislang kaum systematisch untersucht. Insbesondere die Forschung zu Industrieplattformen steht noch am Anfang (Butollo, 2023; Dolata, 2024; Dolata & Radig, 2025; Jovanovic et al., 2022). Die bisherige Konzentration der Plattformforschung auf wenige, wenn auch herausragende und dominante Erfolgsfälle im kommerziellen Internet resultiert in einem Erfolgsbias der etwa Zwischen- und Entwicklungsschritte ausblendet (Seibt, 2023). Begegnen lässt sich diesem analytischen Problem mit einer Prozessperspektive auf Plattformen, die versucht die Durchdringung von Plattformen in Infrastrukturen, Wirtschaftsprozessen und (staatlichen) Rahmenbedingungen nachzuvollziehen (Köchling, 2025; Poell et al., 2019). Am Fall industrieller Plattformisierung lässt sich ein solcher, noch unabgeschlossener und ungefestigter Durchdringungsprozess beobachten. In dieser Situation stellen sich Fragen nach Formen und Möglichkeiten der Governance umso dringender.

Für meine Untersuchung dieses Feldes steht die Technisierung pluraler gesellschaftlicher Anforderungen über Plattformen im Vordergrund. Als analytischer Rahmen dient das organisationssoziologische Konzept der Re-Kombination (Besio & Meyer, 2020), das Strategien von Organisationen beschreibt, mit heterogenen und widersprüchlichen Anforderungen umzugehen. Eine zentrale Strategie ist dabei die Übersetzung: Widersprüchliche Logiken werden so formuliert, dass ihre Gemeinsamkeiten betont und Konflikte heruntergespielt werden. Ich erweitere diese Perspektive mit Blick auf den Einfluss digitaler Technologien.
Zentral für die Gestaltung von Plattformen sind technischne Schnittstellen (Interfaces), die der Übertragung und Vernetzung von Daten und Plattformteilnehmer:innen dienen. Insbesondere Plattformregeln und -normen werden über mediale Interfaces vermittelt (Ametowobla & Kirchner, 2023; Dolata & Schrape, 2022). Dabei kann soziale Ordnung auf Plattformen nicht als Ergebnis unmittelbarer Steuerung verstanden werden, sondern als emergenter Effekt im Zusammenspiel verschiedener Elemente (Ametowobla & Kirchner, 2023).
Aus diesem Verständnis lässt sich fragen: Welche Rolle spielen Interfaces für die Verarbeitung von gesellschaftlichen Anforderungen mit Industrieplattformen und Datenökosystemen? Wie wirkt sich die Gestaltung und Nutzung von solchen Interfaces aus auf das komplexe Zusammenspiel von Regeln, Entscheidungsstrukturen, Koordinationsmechanismen und Praktiken (kurz: auf Governance)? Wie ermöglichen und begrenzen spezifische Plattforminterfaces die digitale Transformation von Kooperation und Wettbewerb zwischen Industrieorganisationen?

Interfaces sind in diesem Zusammenhang selbst Gegenstand von Gestaltungs- und Arbeitsprozessen. Das Konzept *Interfacing* (Lipp & Dickel, 2023) dynamisiert die Betrachtungsweise und beschreibt Prozesse der graduellen Verfügbarmachung von sozialen und technischen Faktoren füreinander, etwa technischen, organisationalen und kognitiven Bedingungen in der Gestaltung von User Interfaces und Dashboards. Hierdurch geraten nicht nur Technologien sondern auch die beteiligten Gruppen und ihre Praktiken, sowie deren Rahmenbedingungen wie Vorannahmen, Machtverhältnisse und sozio-technische Zukunftserwartungen (Beckert, 2018; Meyer, 2020; van Lente, 2012) bei der Gestaltung und Verbreitung von Plattforminterfaces in den Blick der Analyse.

Der Beitrag verfolgt die These, die Re-Kombination von gesellschaftlichen Ansprüchen mit sozio-technischen Infrastrukturen wie Plattformen als einen Interfacingprozess zu interpretieren. Hier spielen neben User Interfaces auch technische Schnittstellen eine wesentliche Rolle, die standardisierte Möglichkeiten des Datenaustauschs herstellen.
Empirische Grundlage meines Beitrags ist die empirische Analyse von Expert:inneninterviews, Dokumenten sowie Beobachtungsdaten aus laufenden Technikentwicklungsprojekten mit den Methoden der qualitativen Sozialforschung. Im Zentrum stehen zwei konkrete Fälle industrieller Plattformisierung, die unterschiedliche Gestaltungsansprüche und Verarbeitungsweisen gesellschaftlicher Anforderungen verkörpern: (1) Datenräume (wie *Manufacturing-X* oder *Catena-X*) werden als offene, kooperative Infrastrukturmodelle inszeniert, die durch Standardisierung und geteilte Datenhoheit digitale Souveränität ermöglichen sollen. (2) IIoT-Plattformen (*Siemens Xcelerator*, *Bosch IoT-Suite* oder das mittlerweile eingestellte *ADAMOS*) zielen auf die Integration von Maschinen- und Sensordaten ab, um Echtzeit-Steuerung und neue Dienstleistungen zu ermöglichen.

Die Analyse der Fälle zeigt, dass Governance digitaler Transformation durch spezifische Interfacegestaltung geprägt ist, die Übersetzungsprozesse strukturieren sollen. In Datenräumen werden etwa formelle Standards als Governance-Instrumente genutzt, um offene und interoperable Systeme zu ermöglichen, etwa durch die Definition gemeinsamer Datenmodelle oder Rahmenbedingungen für den Datenaustausch. Aus dieser Offenheit ergibt sich das Folgeproblem einer hohen technischen Komplexität, mit der die Beteiligten umgehen müssen. Dies ist insbesondere bei Versuchen der Überzeugungsarbeit (Lutz & Mölders, 2025; Mölders, 2024) beobachtbar, um potenzielle Teilnehmer für das Datenökosystem zu gewinnen.

IIoT-Plattformen setzen dagegen eher auf proprietäre Steuerungsmechanismen, wie exklusive API-Schnittstellen oder Lizenzmodelle, die zwar die Integration von Daten erleichtern, aber gleichzeitig Abhängigkeiten schaffen und somit Offenheit und Flexibilität vermissen lassen. Gescheiterte Plattformprojekte dokumentieren die Ambiguitäten und Herausforderungen, die mit diesem Modus der Plattformisierung einhergehen.
Der Beitrag adressiert die Governance digitaler Transformation in erster Linie analytisch. Daneben soll der Beitrag auch interdisziplinäre bzw. anwendungsorientierte Relevanz entfalten. Der Beitrag ist interdisziplinär anschlussfähig, da er soziologische Konzepte (Re-Kombination, Governance, Zukunftserwartungen) mit Fragen der Technikgestaltung und organisationalen Transformation verknüpft. Er zeigt auf, wie digitale Technologien nicht nur als Werkzeuge, sondern als Medien von Governance-Prozessen wirken. Gleichzeitig wird die Ambivalenz digitaler Transformation betont: Einerseits ermöglichen Plattformen die Bearbeitung gesellschaftlicher Herausforderungen, andererseits reproduzieren sie bestehende Asymmetrien oder schaffen neue.

Diese Forschung ist Teil des aktuell laufenden Projekts „inteRKombi" (2024-2028) am Institut für Soziologie der Universität Mainz, welches durch das Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR) gefördert wird. Unser Projekt hat das Ziel, das soziologische Verständnis darüber zu erweitern, wie Organisationen Digitalisierungsprozesse steuern und gestalten. In diesen Beitrag fließen erste Ergebnisse aus der empirischen Arbeit ein. Der Beitrag dient dazu, wesentliche Ergebnisse dieses Vorhabens in interdisziplinär anschlussfähiger Form aufzubereiten. Der Beitrag kann auf Deutsch oder Englisch präsentiert werden.

Alaimo, C., & Kallinikos, J. (2017). Computing the everyday: Social media as data platforms. *The Information Society*, *33*(4), 175–191. https://doi.org/10.1080/01972243.2017.1318327
Ametowobla, D., & Kirchner, S. (2023). The Organization of Digital Platforms: The Role of Digital Technology and Architecture for Social Order. *Zeitschrift Für Soziologie*, *52*(2), 143–156. https://doi.org/10.1515/zfsoz-2023-2012
Beckert, J. (2018). *Imaginierte Zukunft: Fiktionale Erwartungen und die Dynamik des Kapitalismus* (S. Gebauer, Übers.). Suhrkamp.
Besio, C., & Meyer, U. (2020). Organisationale Re-Kombination und sozialer Wandel : Eine Forschungsagenda. *Soziale Systeme*, *25*(2), 436–456. https://doi.org/10.1515/sosys-2020-0023
Butollo, F. (2023). Das industrielle Internet als Ort der Wertschöpfung und -aneignung: Erobert der Plattformkapitalismus die Industrie? In T. Carstensen, S. Schaupp, & S. Sevignani (Hrsg.), *Theorien des digitalen Kapitalismus: Arbeit, Ökonomie, Politik und Subjekt* (S. 121–140). Suhrkamp.
Dolata, U. (2024). Industrieplattformen als Markt-, Produktions- und Innovationsflächen. Feldvermessungen und theoretisch-konzeptionelle Überlegungen. *Berliner Journal für Soziologie*, *34*(2), 171–196. https://doi.org/10.1007/s11609-024-00526-3
Dolata, U., & Radig, A.-K. (2025). Von Netzwerken zu Plattformen?: Strukturelle Veränderungen in der Organisation industrieller Produktions- und Innovationsprozesse. *Arbeit*, *34*(4), 379–400. https://doi.org/10.1515/arbeit-2025-0019
Dolata, U., & Schrape, J.-F. (2022). Plattform-Architekturen. Strukturation und Koordination von Plattformunternehmen im Internet. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *74*(1), 11–34. https://doi.org/10.1007/s11577-022-00826-7

Dolata, U., & Schrape, J.-F. (2025). Platform companies on the internet as a new organizational form. A sociological perspective. *Innovation: The European Journal of Social Science Research*, *38*(3), 1228–1247. https://doi.org/10.1080/13511610.2023.2182217

Gawer, A. (2022). Digital platforms and ecosystems: Remarks on the dominant organizational forms of the digital age. *Innovation*, *24*(1), 110–124. https://doi.org/10.1080/14479338.2021.1965888

Helmond, A., Nieborg, D. B., & Van Der Vlist, F. N. (2019). Facebook's evolution: Development of a platform-as-infrastructure. *Internet Histories*, *3*(2), 123–146. https://doi.org/10.1080/24701475.2019.1593667

Jovanovic, M., Sjödin, D., & Parida, V. (2022). Co-evolution of platform architecture, platform services, and platform governance*:* Expanding the platform value of industrial digital platforms. *Technovation*, *118*(102218). https://doi.org/10.1016/j.technovation.2020.102218

Kenney, M., & Zysman, J. (2016). The Rise of the Platform Economy. *Issues in science and technology*, *32*(3), 61–69.

Kirchner, S., & Beyer, J. (2016). Die Plattformlogik als digitale Marktordnung: Wie die Digitalisierung Kopplungen von Unternehmen löst und Märkte transformiert. *Zeitschrift für Soziologie*, *45*(5), 324–339. https://doi.org/10.1515/zfsoz-2015-1019

Köchling, S. (2025). Platformization of Scholarly Communication: Academic Social Networking Sites as New Infrastructures and Organizations. *Emerging Media*, *3*(2), 318–342. https://doi.org/10.1177/27523543251346565

Lipp, B., & Dickel, S. (2023). Interfacing the human/machine. *Distinktion: Journal of Social Theory*, *24*(3), 425–443. https://doi.org/10.1080/1600910X.2021.2012709

Lutz, K., & Mölders, M. (2025). Organisationale Übersetzung des Unverfügbaren. *Soziologie und Nachhaltigkeit*, *11*(2), 67–85. https://doi.org/10.17879/sun-2025-9154

Meyer, U. (2020). Industrie 4.0 als sozio-technische Zukunftsvorstellung. Zur Bedeutung von organisationaler Sinnerzeugung und -stiftung. In S. Maasen & J.-H. Passoth (Hrsg.), *Soziologie des Digitalen—Digitale Soziologie?: Soziale Welt—Sonderband 23* (1. Aufl., S. 349–378). Nomos. https://doi.org/10.5771/9783845295008-349

Mölders, M. (2024). Die differenzierte Gesellschaft und ihre Probleme: Zur Rolle von Übersetzungsagenturen als Lösungstreiberinnen. *Zeitschrift für Theoretische Soziologie*, 198–223. https://doi.org/10.17879/zts-2024-9088

Nieborg, D., Poell, T., & Van Dijck, J. (2022). Platforms and Platformization. In T. Flew, J. Holt, & J. Thomas (Hrsg.), *The SAGE Handbook of the Digital Media Economy* (S. 29–49). SAGE.

Poell, T., Nieborg, D., & Van Dijck, J. (2019). Platformisation. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1425

Schaupp, S. (2021). *Technopolitik von unten: Algorithmische Arbeitssteuerung und kybernetische Proletarisierung* (Erste Auflage). Matthes & Seitz Berlin.

Seibt, D. (2023). Plattformisierung als Transformation Strategischer Handlungsfelder: Der Fall der Orthopädietechnik. In *Polarisierte Welten. Verhandlungen des 41. Kongresses der Deutschen Gesellschaft für Soziologie 2022*.

Staab, P. (2019). *Digitaler Kapitalismus: Markt und Herrschaft in der Ökonomie der Unknappheit* (Erste Auflage, Originalausgabe). Suhrkamp.

van Dijck, J., Poell, T., & de Waal, M. (2018). *The Platform Society*. Oxford University Press. https://doi.org/10.1093/oso/9780190889760.001.0001

van Lente, H. (2012). Navigating foresight in a sea of expectations: Lessons from the sociology of expectations. *Technology Analysis & Strategic Management*, *24*(8), 769–782. https://doi.org/10.1080/09537325.2012.715478

Zuboff, S. (2018). *Das Zeitalter des Überwachungskapitalismus* (B. Schmid, Übers.). Campus Verlag.

Digitale Souveränität: Die nächste Herausforderung der digitalen Transformation

Walter Brenner | Universität St. Gallen, Schweiz; Barbara Brenner | Universität St. Gallen, Schweiz; Benjamin van Giffen | Universität Liechtenstein, Liechtenstein

Digitale Transformation als intensiver Einsatz der Potentiale der Informations- und Kommunikationstechnik ist eine zentrale Herausforderungen der Gegenwart für Wirtschaft und Gesellschaft. Die Nutzung der Informations- und Kommunikationstechnik in Unternehmen begann bereits in den 60er Jahren des vergangenen Jahrhunderts. Vor ungefähr 15 Jahren ist der Begriff digitale Transformation für die gesamthafte Veränderung von Unternehmen geprägt worden (vgl. Brynjolfsson/McAfee 2014; Hess und Graf 2026).
Im Verlauf der digitalen Transformation stehen Unternehmen vor zahlreichen Herausforderungen und es sind teilweise sehr schwierige Entscheidungen zu treffen. Derzeit stehen die Nutzung der Potentiale Künstlicher Intelligenz, insbesondere von generischer und agentischer Künstlicher Intelligenz sowie die Stärkung der digitalen Souveränität im Vordergrund, wenn man Entscheider in Unternehmen, vor allem Chief Information Officer (CIO) fragt.
Vor diesem Hintergrund beschäftigt sich unser Vortrag mit folgender zentraler Forschungsfrage: Wie steht der Chief Information Officer zu der Forderung nach Stärkung der digitalen Souveränität und welchen Einfluss hat die Forderung nach erhöhter digitaler Souveränität auf ausgewählte Entscheidungsprozesse von Unternehmen und öffentlichen Verwaltungen, insbesondere der Chief Information Officer, im Rahmen der digitalen Transformation?
Digitale Souveränität verstehen wir dabei als die Fähigkeit von Einzelpersonen, Institutionen und staatlichen Organen, in der vernetzten Welt eigenständig und selbstbestimmt zu agieren (Kagermann et al. 2021, Fleming 2025).

Der Vortag basiert auf Forschung, die an der Universität St. Gallen und an der Universität Liechtenstein in den vergangenen drei Jahren gemacht wurde. Dabei wurden zwei Befragungen von CIOs, Workshops und zahlreiche Hintergrundgespräche geführt.

Der Vortrag wird auf ausgewählte Erkenntnisse aus unseren Forschungsarbeiten eingehen:

- Das Verbessern der digitalen Souveränität wird als weiteres Kriterium bei Entscheidungen im Rahmen der digitalen Transformation, insbesondere bei der Auswahl von Lieferanten und bei Entscheidungen über die Auswahl von Hard- und Software eine Rolle spielen. Dabei besteht ein Zielkonflikt zwischen der Forderung nach Reduktion von Kosten in der IT und der Erhöhung der digitalen Souveränität.
- CIOs sind mit dem Verhalten grosser Softwareanbieter nicht sehr zufrieden. Sie wünschen sich, dass sie jede einzelne Komponente von Softwarepakten einzeln bestellen können (84%). CIOs sind mit der «Bundlingpraxis» grosser Softwareanbieter sehr unzufrieden (77%). CIOs wollen früher über Veränderungen der Software informiert werden (100%).
- Die Abhängigkeiten von grossen amerikanischen und teilweise chinesischen Anbietern wird durch den starken Einsatz Künstliche Intelligenz noch steigen.
- Die Bereitschaft der CIOs Abstriche beispielsweise bei Funktionalität, Antwortzeiten, Skalierbarkeit oder gar dem Preis in Kauf zu nehmen, um Datensouveränität, operationelle Souveränität oder rechtliche Souveränität zu stärken, ist im Durchschnitt nicht sehr gross (max 20%).
- CIOs sehen im Moment keine realistische Chance, sich aus den Abhängigkeiten von wenigen grossen Anbietern zu befreien.
- Hinweise auf sog. Kill Switches, d.h. das gezielte Abschalten digitaler Dienstleistungen durch Anbieter werden nach Ansicht einiger CIOs überbewertet.
- CIOs aus dem nahen Osten und Asien sind unabhängiger als ihre europäische Kolleginnen und Kollegen, wenn es um digitale Souveränität geht.

Der Vortrag schliesst mit einer kritischen und realistischen Analyse der bisherigen Forschungsarbeiten ab. In wenigen Worten ausgedrückt lautet das Fazit: Digitale Souveränität ist als europäische Herausforderung für eine nachhaltige digitale Transformation erkannt. Es wird aber noch viel Zeit brauchen und viel Geld kosten bis es gelingen kann sich aus den Abhängigkeiten von amerikanischen und chinesischen Anbieter zu befreien und die digitale Souveränität nachhaltig zu stärken.

Literatur

Brynjolfsson, Erik, McAfee, Andrew (2014): The Second Machine Age: Work, progress, and prosperity in a time of brilliant technologies. New York: W. W. Norton & Company.

Fleming, Sean (2025). What is digital sovereignty and how are countries approaching it? World Economic Forum. Online verfügbar unter https://www.weforum.org/stories/2025/01/europe-digital-sovereignty/, zuletzt geprüft am 13. April 2026)

Hess, Thomas; Graf, Angela (2026): Beyond the Digital Transformation? Towards a Meta-Concept for Managing the Digital Change of Companies. In: Journal of Information Technology, online first. DOI: 10.1177/02683962261447507.

Kagermann, Henning, Streibich, Karl-Heinz & Suder, Katrin (2021): Acatech Impulse, Digital Sovereignty-Status Quo und Impulse, Acatech 2021. Online verfügbar unter https://www.acatech.de/publikation/digitale-souveraenitaet-status-quo-und-handlungsfelder/download-pdf/?lang=en, zuletzt geprüft am 13.04.2026.

Governance als Enabler: Menschliche Adaption als zentrale, vernachlässigte Steuerungsdimension digitaler Transformation

Imme Baumüller | Unabhängige Forscherin und Unternehmensberaterin, Digitale Transformation

1. Problemstellung und Forschungslücke

Digitale Transformation gilt als strategischer Imperativ — und scheitert mit bemerkenswerter Regelmäßigkeit. Empirische Befunde zeichnen ein ernüchterndes Bild: Eine auf 825 Senior Managern und 70 Transformationsprojekte gestützte Studie der Boston Consulting Group beziffert die Misserfolgsquote auf 70 Prozent (Forth et al., 2020); McKinsey bestätigt diesen Befund konsistent über mehrere Publikationen (Garcia, 2022). Oludapo et al. (2024) referieren auf Basis mehrerer Primärstudien Misserfolgsquoten von über 80 Prozent.

Eine Analyse von Oludapo et al. (2024) im Journal of Business Research identifiziert vier zentrale Dimensionen des Scheiterns — Technologie, Innovation, Management und Informationssysteme. Besonders aufschlussreich ist, dass bis heute trotz jahrzehntelanger Bemühungen in genau diesen Dimensionen zahlreiche digitale Transformationen scheitern: Die Dimensionen sind offenbar notwendige, aber keine hinreichenden Steuerungsgrößen. Was ihnen fehlt, ist die Ebene menschlicher Adaption — die Frage, unter welchen kognitiven, motivationalen und sozialen Bedingungen Individuen neue Arbeitsweisen tatsächlich und dauerhaft übernehmen. Hanelt et al. (2021) benennen in ihrer systematischen Literaturübersicht von 279 Peer-reviewed-Artikeln explizit Mitarbeiterzufriedenheit, Commitment und organisationale Identifikation als offene Forschungsdesiderate — ein Befund, der diese Lücke auf breiter empirischer Basis bestätigt.
Bestehende Frameworks für die Steuerung digitaler Transformationen — von klassischen ITGovernance-Konzeptionen (Weill & Ross, 2004) bis zu aktuellen Digital Transformation Strategy-Ansätzen (Matt et al., 2015) — adressieren vorrangig strukturelle Steuerungsprobleme: Mandate, Verantwortlichkeiten, Prozessarchitekturen und KPI-Systeme. Wo menschliche Faktoren erscheinen, werden sie als operative Implementierungsherausforderung behandelt, nicht als eigenständige Governancedimension (Matt et al., 2015). Governance wird damit primär als Kontroll- und Complianceinstrument — als Prohibitor — konzipiert. Eine wesentliche Steuerungspotenz bleibt ungenutzt.

Der vorliegende Beitrag argumentiert, dass darin ein kategorialer Fehler liegt. Verhaltens- und kognitionswissenschaftliche Erkenntnisse zu Verhaltensänderung in Organisationen existieren — sie sind jedoch nicht als eigenständige Steuerungsdimension in Governance-Frameworks integriert. Erst Governance, die auch die kognitiven, motivationalen und sozialen Bedingungen von Verhaltensänderung aktiv gestaltet — die also als Enabler operiert —, schließt die Lücke zwischen struktureller Steuerung und gelebter Transformation. Dieser Beitrag entwickelt ein erweitertes Governance-Modell, das die Verhaltensebene als eigenständige, gleichrangige Steuerungsdimension integriert.

2. Theoretische Einbettung

Das Konzept der IT-Governance wurde maßgeblich durch Weill und Ross (2004) geprägt, die es als Gesamtheit von Entscheidungsrechten und Verantwortlichkeiten für IT-bezogene Entscheidungen definieren. Matt et al. (2015) erweitern diesen Rahmen auf digitale Transformationsstrategien und identifizieren vier Kerndimensionen: Technologieeinsatz, Wertschöpfungsveränderung, strukturelle Veränderung und finanzielle Aspekte. Governance wird in diesem Forschungsstrang überwiegend als strukturelles Phänomen konzipiert — als Frage der richtigen Mandate, Prozesse, Kontrollmechanismen und Kennzahlen. Was in diesen Modellen weitgehend fehlt, ist eine explizite Konzeptualisierung der Verhaltensebene: Strukturen determinieren kein Verhalten. Zwischen der Gestaltung von Rahmenbedingungen und der tatsächlichen Verhaltensänderung von Individuen liegt eine Lücke, die weder durch bessere Prozesse noch durch schärfere KPIs geschlossen werden kann.

Die Frage, unter welchen Bedingungen Individuen neue Technologien in Organisationen annehmen und dauerhaft nutzen, ist seit Jahrzehnten intensiv erforscht. Das Technology Acceptance Model (TAM; Davis, 1989) und seine Weiterentwicklungen — darunter UTAUT (Venkatesh et al., 2003) und UTAUT2 (Venkatesh et al., 2016) — haben robuste Prädiktoren für Technologieadoption identifiziert: wahrgenommene Nützlichkeit, Aufwandserwartung, sozialer Einfluss sowie Gewohnheit und hedonische Motivation. Die Organisationspsychologie ergänzt diesen Befund: Oreg et al. (2011) zeigen in einer 60-jährigen Literaturübersicht, dass Reaktionen auf Wandel dreidimensional sind — affektiv, kognitiv und behavioral — und dass Kommunikationsqualität sowie individuelle Veränderungsbereitschaft starke Prädiktoren für Adoptionserfolg darstellen. Widerstand ist kein Versagen der Belegschaft, sondern ein Indikator für gestaltbare psychologische Prozesse.

Die Verhaltensökonomie liefert präzise Erklärungen dafür, warum Verhaltensänderung auch unter günstigen Rahmenbedingungen scheitert: Milkman (2021) identifiziert individuell variierende Barrieren wie Prokrastination und Impulsivität, die einheitliche Interventionsstrategien — etwa eine unternehmensweite Kommunikationskampagne oder ein einmaliges Training — systematisch nicht adressieren können. Oettingen (2014) belegt experimentell, dass reine Zielkommunikation die Wahrscheinlichkeit der Zielerreichung sogar senken kann. Thaler und Sunstein (2008) zeigen, dass Entscheidungsarchitekturen Verhalten vorhersagbar beeinflussen — ohne Zwang oder Anreize.

Dieser Wissensbestand ist vorhanden — er ist jedoch in den Governance-Frameworks für digitale Transformation nicht verankert. Governance-Frameworks operieren weiterhin primär auf der strukturellen Ebene und werden — sowohl in der Forschung als auch in der Praxis — vorrangig als Kontroll- und Complianceinstrument verstanden, nicht als Enabler von Verhaltensänderung. Aus über einem Jahrzehnt Praxiserfahrung im Aufbau und der Skalierung digitaler Organisationen lässt sich beobachten, dass verhaltenspsychologische Erkenntnisse in der operativen Ausgestaltung von Governance-Modellen kaum systematisch berücksichtigt werden. Erfolgreiche Unternehmen investieren dabei mindestens ebenso viel in Transformationsmanagement und Adoption wie in die technische Entwicklung — ein Verhältnis, das als Faustregel aus der Praxis Bestand hat, in der Governance-Literatur jedoch keine konzeptionelle Entsprechung findet (aus der eigenen Transformationspraxis der Autorin).

Der vorliegende Beitrag setzt genau an dieser Schnittstelle an: nicht als Beitrag zur Verhaltenspsychologie, sondern als konzeptionelle Erweiterung von Governance-Modellen — mit dem Ziel, Governance vom Prohibitor zum Enabler zu machen.

3. Konzeptuelles Argument: Ein erweitertes Governance-Modell

Aktuelle Governance-Modelle für digitale Transformation operieren primär auf der Ebene formaler Strukturen. Sie regeln, wer welche Entscheidungen trifft und Zugriffe hat, welche Prozesse gelten und wie Erfolg gemessen wird. Was sie nicht regeln, ist die Frage, unter welchen Bedingungen Menschen diese Strukturen tatsächlich mit Leben füllen — also ihr Verhalten dauerhaft verändern. Verhaltensänderung folgt anderen Logiken als Strukturveränderung: Sie erfordert Zeit, Wiederholung, soziale Einbettung und individuelle Passung. Governance-Systeme, die dies ignorieren, produzieren das in der Praxis vielfach beobachtbare Phänomen: Strukturen sind vorhanden, Technologien sind implementiert, Trainings wurden durchgeführt — und dennoch verändert sich die tatsächliche Arbeitsweise nicht.

Der vorliegende Beitrag schlägt ein erweitertes Governance-Modell vor, das drei gleichrangige Ebenen integriert:

Ebene 1 — Strukturelle Governance
Umfasst die in der Literatur etablierten Elemente: Mandate, Entscheidungsrechte, Prozessarchitekturen, KPI-Systeme und die organisatorische Verankerung digitaler Führung. Diese Ebene ist notwendige, aber nicht hinreichende Bedingung. Strukturen schaffen Rahmenbedingungen — sie determinieren kein Verhalten.
Aus der eigenen Transformationspraxis der Autorin: Eine crossfunktionale Digital- und Dateneinheit, die ohne klares Mandat und ohne direkten Zugang zur Geschäftsführungsebene eingerichtet wurde, blieb trotz hoher fachlicher Kompetenz strukturell isoliert. Behaviorale und kommunikative Interventionen konnten dieses strukturelle Defizit nicht kompensieren — und umgekehrt erzeugte das fehlende Mandat allein keine Verhaltensänderung in den Fachbereichen.

Ebene 2 — Kognitiv-kommunikative Governance
Adressiert, wie Transformationsziele gerahmt, kommuniziert und in organisationale Bedeutungsstrukturen eingebettet werden. Kommunikation ist hier nicht als Informationsübermittlung zu verstehen, sondern als Überzeugungsprozess: Welche Narrative machen Veränderung für unterschiedliche Stakeholder bedeutsam? Wie werden Widerstände als Information genutzt statt ignoriert? Diese Ebene zieht Einsichten aus der Persuasionsforschung (Cialdini, 2001) und der Kommunikationswissenschaft heran.
Aus der eigenen Transformationspraxis der Autorin: Die Einführung eines datenbasierten Entscheidungsunterstützungstools scheiterte nicht an der methodischen Qualität, sondern an fehlender kommunikativer Einbettung. Die Ergebnisse wurden der Entscheidungsrunde als analytischer Befund präsentiert — ohne Rückkopplung an die Ziele und Sprache der Empfänger. Die Entscheidung fiel auf Basis von Intuition und Erfahrung. Wer keine gemeinsame Bedeutungsstruktur mit Stakeholdern herstellt, erzeugt keine Adoption — unabhängig von der Qualität der Lösung.

Ebene 3 — Behaviorale Governance
Operationalisiert, durch welche Mechanismen das tatsächliche Alltagsverhalten von Mitarbeiterinnen und Mitarbeitern beeinflusst wird: Anreizstrukturen, die neue Verhaltensweisen sichtbar machen und belohnen; Entscheidungsarchitekturen, die erwünschtes Verhalten zur einfacheren Option machen (Thaler & Sunstein, 2008); soziale Mechanismen wie Peer-Vergleiche; sowie individuell angepasste Interventionen gegen heterogene Adoptionsbarrieren (Milkman, 2021). Behaviorale Governance arbeitet nicht mit Appellen, sondern mit Systemgestaltung.
Aus der eigenen Transformationspraxis der Autorin: Bei der unternehmensweiten Einführung einer Datenplattform erzeugte die Einbettung adoptionsbezogener KPIs — Nutzungsraten, Self-Service- Anfragen, Datenqualitätsbeiträge — in den regulären Führungsreport sowie die partielle Kopplung von Bereichsleiter-Boni an Adoptionsziele erst die nachhaltige Verhaltensänderung. Umfangreiche Trainings und technisch einwandfreie Infrastruktur allein hatten dies nicht geleistet.
Das Modell ist explizit nicht hierarchisch: Die drei Ebenen sind interdependent. Strukturelle Governance ohne behaviorale Fundierung bleibt wirkungslos; behaviorale Interventionen ohne strukturelle Absicherung verpuffen; kommunikative Rahmung ohne strukturelle Rückendeckung erzeugt Zynismus. Erst im Zusammenspiel entsteht nachhaltige Adoption.

Die zentrale konzeptuelle Konsequenz: Die Verhaltensebene muss von einem Implementierungsproblem zu einem Governanceproblem aufgewertet werden. Wenn Verhaltens-Governance als eigenständige Dimension anerkannt wird, hat das Konsequenzen für Zuständigkeiten, Ressourcenallokation, Erfolgsmessung und die Kompetenzprofile digitaler Führungskräfte. Nicht fehlende Strukturen allein, sondern die Abwesenheit einer behavioralen Steuerungslogik ist eine eigenständige, strukturell vernachlässigte Ursachenkategorie für das Scheitern digitaler Transformationen.

4. Methodische Vorgehensweise
Der Beitrag ist konzeptuell-theoretischer Natur. Er kombiniert eine systematische Sichtung der Literatur an der Schnittstelle von Governance-Forschung, Technologieakzeptanzforschung, Organisationspsychologie und Verhaltensökonomie mit illustrierenden Praxisbeispielen aus der eigenen Transformationstätigkeit der Autorin. Die Literaturarbeit integriert Erkenntnisse aus vier Disziplinen: Informationssystems- und Managementforschung, Organisationspsychologie, Verhaltensökonomie sowie Kommunikationswissenschaft. Sie ist nicht als vollständige systematische Literaturübersicht angelegt, sondern als konzeptionelle Synthese: Es werden Forschungsstränge zusammengeführt, die bislang — trotz ihrer Überlappung im Gegenstandsbereich — nicht systematisch im Rahmen von Governance- Modellen integriert wurden.
Die Praxisbeispiele sind aus der eigenen Berufserfahrung der Autorin — über zehn Jahre Aufbau und Skalierung digitaler, Daten- und KI-Organisationen — abgeleitet, anonymisiert und illustrativ in das konzeptuelle Argument eingebettet. Sie sind keine empirischen Belege, sondern dienen der Konkretisierung der drei Governance-Ebenen. Die hybride Perspektive der Autorin — Dr. phil. In Kommunikationswissenschaft, MBA und operative Führungserfahrung — ist dabei methodischer Vorzug, nicht Limitation.

5. Interdisziplinärer Beitrag und Implikationen
Der Beitrag leistet einen konzeptuellen Beitrag zur Governance-Forschung im Bereich digitaler Transformation in Unternehmen, indem er die Verhaltensebene als eigenständige, gleichrangige Steuerungsdimension einführt und mit bestehenden Governance-Konzepten verknüpft. Das vorgeschlagene Drei-Ebenen-Modell integriert Erkenntnisse aus der Verhaltensökonomie, der Organisationspsychologie und der Kommunikationswissenschaft in einen Governancerahmen — eine Verknüpfung, die in der bisherigen Transformationsforschung nicht systematisch vollzogen wurde. Darüber hinaus trägt der Beitrag zur Konzeptualisierung von Governance-Versagen bei: Nicht nur fehlende Strukturen, sondern die Abwesenheit einer behavioralen Steuerungslogik ist eine eigenständige Ursachenkategorie für das Scheitern digitaler Transformationen.

Der Beitrag schließt an das im Call skizzierte breite Governance-Verständnis an — Governance als Ermöglichung, Begrenzung und Stabilisierung von Transformationsprozessen (Mayntz, 2004) — und erweitert es um ein bislang vernachlässigtes Gestaltungsfeld. Er adressiert insbesondere die Themenfelder „Gestaltungs- und Interventionsansätze zur reflexiven Governance digitaler Transformation“ sowie „Governance digitaler Transformationsprozesse in Unternehmen“. Der Beitrag ist an der Schnittstelle von Managementwissenschaft, Informationssystemforschung, Organisationspsychologie und Kommunikationswissenschaft verortet und eröffnet damit den im Call angestrebten Dialog zwischen disziplinären Zugängen.

Für Organisationen ergeben sich konkrete Handlungsfelder: die systematische Diagnose von Adoptionsbarrieren als Governance-Aufgabe; Kommunikation als gesteuerter Überzeugungsprozess mit differenzierten Narrativen; Anreiz- und Entscheidungsarchitekturen als operative Steuerungsinstrumente; sowie Kompetenzprofile für digitale Führungskräfte, die verhaltenspsychologisches Grundwissen einschließen.

Literaturverzeichnis
Cialdini, R. B. (2001). Influence: Science and practice (4th ed.). Allyn & Bacon.
Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340.
Forth, P., de Laubier, R., Reichert, T., & Chakraborty, S. (2020). Flipping the odds of digital transformation success. Boston Consulting Group. https://www.bcg.com/publications/2020/increasing-odds-of-success-in-digitaltransformation

Garcia, J. (2022). Common pitfalls in transformations: A conversation with Jon Garcia. McKinsey & Company. https://www.mckinsey.com/capabilities/transformation/our-insights/common-pitfalls-in-transformations-aconversation-with-jon-garcia
Hanelt, A., Bohnsack, R., Marz, D., & Antunes Marante, C. (2021). A systematic review of the literature on digital transformation: Insights and implications for strategy and organizational change. Journal of Management Studies, 58(5), 1159–1197. https://doi.org/10.1111/joms.12639
Matt, C., Hess, T., & Benlian, A. (2015). Digital transformation strategies. Business & Information Systems Engineering, 57(5), 339–343. https://doi.org/10.1007/s12599-015-0401-5
Mayntz, R. (2004). Governance Theory als fortentwickelte Steuerungstheorie? MPIfG Working Paper 04/1.
Milkman, K. (2021). How to change: The science of getting from where you are to where you want to be. Portfolio/Penguin.
Oettingen, G. (2014). Rethinking positive thinking: Inside the new science of motivation. Portfolio/Penguin.
Oludapo, S., Carroll, N., & Helfert, M. (2024). Why do so many digital transformations fail? A bibliometric analysis and future research agenda. Journal of Business Research, 174, Article 114528. https://doi.org/10.1016/j.jbusres.2024.114528
Oreg, S., Vakola, M., & Armenakis, A. (2011). Change recipients' reactions to organizational change: A 60-year review of quantitative studies. Journal of Applied Behavioral Science, 47(4), 461–524. https://doi.org/10.1177/0021886310396550
Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Yale University Press.
Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425–478.
Venkatesh, V., Thong, J. Y. L., & Xu, X. (2016). Unified Theory of Acceptance and Use of Technology: A Synthesis and the Road Ahead. Journal of the Association for Information Systems, 17(5), 328–376.
Weill, P., & Ross, J. W. (2004). IT governance: How top performers manage IT decision rights for superior results. Harvard Business School Press.

COMMUNICATION & SOCIETY Session 2
Digital Public Sphere: Conditions and Framework for a Constructive Debating Culture

19.11.2026, 13:15-15:15

Language: English

Programme Committee: Andreas Wenninger | bidt; Hannah Schmid-Petri | bidt & University of Passau; Ute Schmid | bidt & University of Bamberg; Andreas Jungherr | bidt & University of Bamberg

Session Chair: Andreas Wenninger | bidt

Relational Truths and Synthetic Publics: Political Communication in the Age of Generative AI

Ignas Kalpokas | Vytautas Magnus University, Lithuania

Digital information and communication technologies have profoundly reconfigured the conditions under which political communication unfolds. While these technologies have expanded opportunities for participation, visibility, and expression, they have simultaneously destabilised shared epistemic foundations and enabled new forms of manipulation, including disinformation campaigns, coordinated inauthentic behaviour, and algorithmically amplified outrage. The rapid proliferation of generative artificial intelligence (GenAI) introduces an additional layer of complexity by drastically lowering the barriers to content production while rendering distinctions between human- and machine-generated communication increasingly negligible. Against this backdrop, the paper asks how political publics function under conditions shaped by generative AI and what this implies for the production of truth, the dynamics of political communication, and the prospects for democratic debate.

The central argument developed in this paper is that contemporary digital publics cannot be adequately understood through traditional frameworks that assume stable truths, clearly bounded publics, or controllable media environments. Instead, they must be conceptualised as relational, dynamic, and socio-technical assemblages in which truth claims emerge contingently through interactions among human actors, technological systems, and institutional structures. In such a context, the problem is not simply the spread of misinformation but the transformation of the very conditions under which truth is produced, perceived, and contested.

This shift necessitates a rethinking of the notion of ‘post-truth’. Rather than interpreting post-truth as a deviation from an earlier era of factual consensus, it is more productive to understand it as a mode of life in which established distinctions between fact and belief, online and offline, or rationality and emotion become increasingly blurred. Reality itself becomes interaction-contingent, characterised by instability, multiplicity, and continuous renegotiation. Under such conditions, the pursuit of objective truth as a stable endpoint gives way to an analysis of how specific configurations of relations produce particular iterations of truth (Allan 2023; Wehrle 2023). This perspective also highlights that emotional expression and affective communication should not be dismissed as inherently irrational; rather, they constitute central components of contemporary political communication (Hannon 2023). Consequently, the key analytical focus shifts from factual validity alone to political efficacy and the capacity of truth claims to resonate within particular socio-technical environments (Anastasiou 2025).

To make sense of these processes, the paper adopts an assemblage-based understanding of publics. Publics are conceptualised not as unified arenas of deliberation but as dynamic constellations of actors, frames of relevance, and communicative practices (Hasebrink, Merten and Behre 2023). These publics are multiple, overlapping, and continuously in flux, forming a network of interconnected communicative formations rather than a single public sphere (Bruns 2023). Crucially, they are co-constituted by human and non-human elements alike, including algorithms, interfaces, and data infrastructures. The notion of the ‘feed-as-assemblage’ is

particularly instructive here, as it captures how political information is curated through the interplay of recommender systems, user practices, and interface design (Bareither and Wirth 2026). In this sense, political communication is not merely transmitted but actively produced within socio-technical systems.

Generative AI intensifies these dynamics by functioning as an active participant in communication rather than a passive tool. It gives rise to what has been described as 'artificial sociality', wherein machines simulate social interaction, emotional engagement, and communicative intent in ways that elicit meaningful responses from users (Depounti and Natale 2025). This simulation is not inconsequential; on the contrary, it shapes how individuals perceive, interpret, and act upon information. At the same time, generative AI operates through probabilistic processes that introduce a fundamental unpredictability into communication. Its outputs emerge from latent spaces that recombine patterns in ways that cannot be fully anticipated or controlled, thereby challenging traditional assumptions about the governability of media environments (Amoore et al. 2024). Moreover, by privileging statistically dominant patterns, AI systems tend to reinforce existing norms and biases, contributing to what has been described as epistemic conformism (Liu 2025; Miragoli 2025).

Methodologically, the paper adopts a theory-driven and conceptually integrative approach. It synthesises insights from media studies, political communication, and science and technology studies to develop a framework for understanding political communication under conditions of generative AI. The analysis draws on a wide range of studies to identify recurring patterns, tensions, and transformations, with the aim of generating theoretically grounded propositions that can inform future empirical research.

The analysis yields several key insights. First, the proliferation of generative AI contributes not merely to misinformation but to a broader pluralisation of truth claims. The capacity to rapidly generate and customise content enables the creation of highly specific narratives tailored to particular audiences, including forms of 'political fanfiction' that are not necessarily intended to deceive but to express and reinforce group identities (Yeilaghi and Mou 2025). Even when such content is recognised as synthetic, it can still shape perceptions by seeding doubt and undermining trust (Cazzamatta and Sarısakaloğlu 2025). In this sense, the boundary between truth and fiction becomes less relevant than the relational and affective dynamics through which content gains traction.

Second, political communication in AI-mediated environments is increasingly driven by emotional and identity-based appeals. Platform logics that prioritise engagement tend to amplify content that provokes strong reactions, including outrage and indignation. AI-generated 'rage bait' exemplifies how communicative strategies are optimised for visibility rather than informational quality (Cover 2025). At the same time, individuals are more likely to accept and share information that aligns with their existing beliefs and group identities, reinforcing processes of informational and normative conformity (Chadwick and Stanyer 2022; Chan 2024). Under these conditions, the effectiveness of communication depends less on its factual accuracy than on its resonance within specific relational contexts.

Third, generative AI simultaneously democratises and fragments political communication. On the one hand, it enables a wider range of actors to participate in content production, allowing supporters and grassroots communities to generate campaign materials and political narratives without institutional mediation (Yeilaghi and Mou 2025). On the other hand, this proliferation of content contributes to the formation of increasingly segmented publics, each characterised by its own communicative norms and epistemic orientations. Rather than a shared public sphere, what emerges is a fluid and heterogeneous network of micro-publics in which consensus becomes increasingly difficult to achieve.

Fourth, platform infrastructures play a decisive role in structuring these dynamics. Their power lies less in directly controlling content and more in shaping visibility, attention, and amplification (Zeng and Kaye 2022). Through

algorithmic curation, platforms influence which narratives gain prominence and which remain marginal, thereby shaping the contours of public discourse. These processes are further reinforced by economic incentives that prioritise engagement and monetisation, contributing to the proliferation of sensationalist and polarising content (Miklian and Hoelscher 2026). At the same time, users tend to attribute a higher degree of objectivity to algorithmic outputs and exhibit a relative leniency towards algorithmic bias, a phenomenon that complicates efforts to foster critical engagement (Aroyo et al. 2021; Messingschlager and Appel 2025).
Finally, trust in digital publics increasingly emerges as a relational phenomenon rather than a function of institutional authority. In social media environments, credibility is often determined by who shares a piece of content and how it is embedded within social networks, rather than by its source in a traditional sense (Park and Lee 2023). This shift underscores the limitations of approaches that focus solely on improving information quality or strengthening institutional credibility, as they fail to address the underlying relational dynamics of trust formation.

Taken together, these findings suggest that the challenges facing democratic communication in the age of generative AI cannot be resolved through simple solutions such as fact-checking or content moderation alone. Instead, they require a more fundamental rethinking of the design and governance of digital environments. Interfaces should be conceptualised not merely as neutral conduits but as political sites that shape perception, attention, and interaction (Lipp et al. 2025). Enhancing transparency and user control is therefore essential, but it must be accompanied by measures that address the broader structural incentives driving engagement-based amplification. Likewise, the acceleration and fragmentation of digital communication often leave little room for deliberation, contributing to what has been described as a failure to process and critically engage with information (Almeida-Amir 2025). Addressing this issue requires both design interventions and broader regulatory frameworks that prioritise public values over purely commercial objectives (Graef and Bostoen 2025).

In conclusion, the transformation of political communication in the age of generative AI is best understood as a shift from stable, shared truths to a landscape characterised by relationally produced and politically contested truth claims. Publics are no longer bounded arenas of deliberation but dynamic assemblages in which human and artificial agents jointly produce meaning. Navigating this landscape requires moving beyond binary distinctions and embracing a process-oriented perspective that foregrounds the complex interplay between technology, society, and politics. Only on this basis can effective strategies be developed to sustain trust, participation, and democratic debate in digitally mediated public spheres.

This presentation contributes to interdisciplinary debates on digital public spheres by linking political communication, media studies, and science and technology studies to conceptualise publics as relational, socio-technical assemblages shaped by generative AI. It moves beyond viewing publics as unified arenas and instead emphasises their multiplicity, fluidity, and co-constitution by human actors, algorithms, and platform infrastructures. By shifting the analytical focus from fixed binaries (e.g. truth vs. falsehood) to the relational emergence of meaning and credibility, it offers a shared framework for examining how technological, social, and institutional dynamics jointly structure public discourse.

At the same time, the paper identifies key prerequisites for a constructive culture of debate under AI-mediated conditions. It argues that such a culture depends less on content moderation alone than on the structural conditions of communication, including transparent and reflexive platform design, regulatory frameworks addressing engagement-driven incentives, and environments that enable critical reflection rather than reactive communication. It further emphasises the need to integrate affect and identity into deliberative models and to strengthen trust as a relational achievement. Together, these elements outline a framework in which constructive debate emerges from the interplay of technological design, social practice, and governance.

Bibliography

Allan, Keith (2023). Why Truth Is Necessarily Pragmatic. Intercultural Pragmatics, 20(3): 251–267.
Almeida-Amir, Maia (2025). Obedient Attention: Doomscrolling and the Production of Authoritarian Subjectivity. Dialogues on Digital Society, 1(3): 438-441.
Amoore, Louise et al. (2024). A World Model: On the Political Logics of Generative AI. Political Geography, 113: 1-9.
Anastasiou, Michaelangelo (2025). The Hegemonic World Picture: Representation, Post-truth, and Artificial Intelligence. Philosophy & Social Criticism, doi: 10.1177/01914537251374951.
Aroyo, Alexander M. et al. (2021). Overtrusting Robots: Setting a Research Agenda to Mitigate Overtrust in Automation. Paladyn, Journal of Behavioral Robotics, 12: 423-436.
Bareither, Christoph and Wirth, Sabine (2026). Social Media Feeds as Curatorial Assemblages: A Conceptual Framework. Media, Culture & Society, 48(1): 73-88.
Bruns, Axel (2023). From 'the' Public Sphere to a Network of Publics: Towards an Empirically Founded Moel of Contemporary Public Communication Spaces. Communication Theory, 33: 70-81.
Cazzamatta, Regina and Sarısakaloğlu, Ainur (2025). AI-Generated Misinformation: A Case Study on Emerging Trends in Fact-Checking Practices Across Brazil, Germany, and the United Kingdom. Emerging Media, 3(2): 214-251.
Chadwick, Andrew and Stanyer, James (2022). Deception as a Bridging Concept in the Study of Disinformation, Misinformation, and Misperceptions: Toward a Holistic Framework. Communication Theory, 32: 1-24.
Chan, Jovy (2024). Online Astroturfing: A Problem Beyond Disinformation. Philosophy and Social Criticism, 50(3): 507-528.
Cover, Rob (2025). AI Generation of Rage Bait: Implications for Digital Harms. New Media & Society, doi: 10.1177/14614448251400675.
Depounti, Iliana and Natale, Simone (2025). Decoding Artificial Sociality: Technologies, Dynamics, Implications. New Media & Society, 27(10): 5457-5470.
Graef, Inge and Bostoen, Friso (2025). A Typology of Platform Power and Its Regulation. Information, Communication & Society, doi: 10.1080/1369118X.2025.2512972.
Hannon, Michael (2023). The Politics of Post-Truth. Critical Review: A Journal of Politics and Society, 35(1-2): 40-62.
Hasebrink, Uwe, Marten, Lisa and Behre, Julia (2023). Public Connection Repertoires and Communicative Figurations of Publics: Conceptualizing Individuals' Contribution to Public Spheres. Communicative Theory, 22: 82-91.
Lipp, Benjamin et al. (2025). De-Centring the Interface: Towards the Integrated Study of Interfacial Relations. Information, Communication & Society, 28(8): 1463-1479.
Liu, Zhaoming (2025). Cultural Bias in Large Language Models: A Comprehensive Analysis and Mitigation Strategies. Journal of Transcultural Communication, 3(2): 224-244.
Messingschlager, Tanja V. and Appel, Markus (2025). Algorithmic Bias in Image-Generating Artificial Intelligence: Prevalence and User Perceptions. Information, Communication & Society, doi: 10.1080/1369118X.2025.2584146.
Miklian, Jason and Hoelscher, Kristian (2026). A New Digital Divide? Coder Worldviews, the 'Slop Economy,' and Democracy in the Age of AI. Information, Communication & Society, doi: 10.1080/1369118X.2025.2566814.
Miragoli, Martin (2025). Conformism, Ignorance & Injustice: AI as a Tool of Epistemic Oppression. Episteme: 22: 522–540.
Park, Sora and Lee, Jee Young (2023). Incidental News Exposure on Facebook and its Relation to Trust in the News. Social Media + Society, doi: 10.1177/20563051231158823.
Wehrle, Maren (2023). Can the 'Real World' Please Stand Up? The Struggle for Normality as a Claim to Reality. Philosophy and Social Criticism, 49(2): 151-163.

Yeilaghi, Noureddin and Mou, Yi (2025). Artificial Intelligence and Digital Deification: Examining the Construction of Donald Trump's Superhuman Persona. Javnost – The Public, 32(1): 115-139.
Zheng, Jing and Kaye, D. Bondy Valdovinos (2022). From Content Moderation to Visibility Moderation: A Case Study of Platform Governance on TikTok. Policy & Internet, 14(1): 79-95.

When Transparency Teaches: Explainable AI, Social Learning, and Evasion in Automated Content Moderation

Malika Talgatova | TU Munich; Yannis Theocharis | TU Munich; Nils Weidmann | University of Konstanz; Molly Roberts | University of California San Diego, USA

Digital public spheres depend on moderation systems that can reduce harassment, defamation, manipulation, and other forms of dysfunctional participation without unduly suppressing legitimate expression. This balance has become more difficult as online communication is increasingly governed by automated classifiers, large-scale platform rules, and, more recently, generative AI systems that make the production and variation of content cheaper (Gorwa, Binns and Katzenbach, 2020; Calatayud, 2025; Instagram Help Center, n.d.; Oversight Board, 2024; Lawler, 2026). Calls for transparency and explainable AI in platform governance are therefore normatively compelling: users should know why their speech is restricted, platforms should be accountable, and public oversight requires some visibility into how rules are enforced (Suzor et al., 2019; Suzor, 2019; Jhaver, Bruckman and Gilbert, 2019; Urman and Makhortykh, 2023, etc.). Yet algorithmic transparency also creates a behavioral problem (Br¨uckner, Kanzow and Scheffer, 2012; Hardt et al., 2016; Bambauer and Zarsky, 2018; Ghalme et al., 2021; Wang et al., 2023, etc.). In content moderation as well, where users are strategic, explanations may not only help people comply with rules; they may also help them preserve prohibited meaning while changing surface wording so that content passes automated enforcement. This is the reasoning behind opacity adopted by many social platforms as well, like for example, Meta or Google (Zuckerberg, 2024; Google, 2024).

This project asks: How do different forms of moderation transparency and social information shape users' ability to understand, comply with, and strategically work around an automated content moderation system? The study focuses on a central dilemma for digital public spheres: the same informational tools that make moderation more accountable may also create learning environments in which users infer the boundaries of enforcement. The question is not whether transparency is "good" or "bad" in the abstract. Rather, the project distinguishes between different kinds of transparency and different sources of information. It asks whether local model feedback, broader policy-level guidance, and exposure to other users' apparently successful posts have different consequences for constructive participation and moderation evasion.

The conceptual framework combines three literatures that are often discussed separately. First, work on platform governance and content moderation emphasizes that platforms are not neutral intermediaries but institutional actors that structure visibility, participation, and public discourse through rules, interfaces, and enforcement practices (Gillespie, 2018). Second, research on algorithmic transparency and explainable AI shows why explanations matter for accountability, contestability, trust, and human oversight (Cai, Jongejan and Holbrook, 2019; Jhaver et al., 2019; He, Hong and Raghu, 2025; Leibmann et al., 2025). In the content moderation setting, however, explanations are more than post-hoc justifications, as they also provide actionable information for future behavior. Third, strategic classification and strategic machine learning provide a formal vocabulary for this problem. In strategic classification, people respond to classifiers by modifying observable features in order to obtain a desired classification outcome (Br¨uckner, Kanzow and Scheffer, 2012; Hardt et al., 2016; Bj¨orkegren, Blumenstock and Knight, N.d.; Shao, Xie and Yang, 2025). Applied to moderation, the classifier is the automated moderator, the observable feature vector is the text of the post, and users may alter wording while attempting to preserve the underlying message. The "strategic classification in the dark"

perspective (Hardt et al., 2016; Ghalme et al., 2021; Bechavod et al., 2022) is especially relevant: even when the classifier is opaque, users may learn from samples, feedback, and other people's experiences.

The project therefore treats moderation not as a one-shot act of removal but as an interactive system, in which users learn from the platform and from one another. Even a "pass" vs. "block" verdict can teach users something across repeated attempts; a score, threshold, or local explanation can reveal which words are risky; and visible posts that passed moderation can expose the public residue of enforcement, even when the platform reveals little. This shifts the focus from transparency as disclosure to transparency as a behavioral intervention. The key empirical question is whether different informational environments support rule understanding and constructive compliance, or instead facilitate evasion: content that passes the automated system while preserving policy-violating meaning.

Methodological approach and study design: Methodologically, the project uses a preregistered online experiment built around a controlled "captioning game" on a custom-built platform with a single moderation policy and an automated moderator. Participants are shown short captions involving fictional characters and are asked to rewrite each caption so that it passes an automated moderator while preserving as much of the original meaning as possible. The moderation rule prohibits insults or derogatory references to characters' appearances. The use of fictional characters and stylized captions allows the study to examine strategic adaptation to moderation in a lower-risk setting, without targeting real private individuals. At the same time, the task captures a common governance problem on social media: users may alter wording, phrasing, or indirectness while attempting to keep the communicative harmful intent intact.

The design crosses two treatment dimensions. The first dimension varies moderation transparency. In the baseline condition, participants receive only the verdict: whether their post passes or fails the automated moderator. In the local model feedback condition, participants see the verdict, model score, moderation threshold, and local feedback about which words or phrases appear most responsible for the model's "block" decision. This condition operationalizes a narrow, post-specific form of explainable AI: it gives users tactical information about the model's decision on their own submission. In the policy/global feedback condition, participants see the verdict, score, and threshold, plus broader guidance about the types of wording patterns the moderation system was trained to detect. This condition operationalizes a more general, policy-like form of transparency: it gives users information about the moderation regime rather than only about a single post.

The second dimension varies social information. In the no-feed condition, participants do not see examples from other users. In the non-offensive feed condition, participants see posts that passed the moderator and do not violate the underlying policy. This condition captures exposure to ordinary allowed content: it shows what is permitted without directly demonstrating evasion. In the mixed-feed condition, participants see a mixture of posts that passed the moderator, including both non-offensive posts and offensive or evasive posts that passed despite violating the policy. This condition captures a more realistic public feed, where users encounter not only compliant examples but also borderline or harmful content that appears to have survived enforcement. The comparison between no feed and non-offensive feed estimates the effect of exposure to passing content in general; the comparison between non-offensive and mixed feeds estimates the added effect of exposure to evasive examples, holding constant the presence of a feed; and the comparison between no feed and mixed feed estimates the total effect of a realistic informational environment.

The primary outcomes combine automated and human evaluation. First, the automated moderator records whether a submitted caption passes or fails. Second, independent human raters judge whether each submitted caption violates the underlying moderation policy. Third, separate pairwise human ratings assess semantic fidelity: whether a rewritten caption preserves the meaning of the original caption compared with other rewrites of the same prompt. The main evasion outcome is defined as automated passage multiplied by human-rated

policy violation. The main high-fidelity evasion outcome further multiplies this by a within-caption semantic fidelity percentile. This measurement strategy is important because passage through the model alone is not evidence of meaningful evasion: a participant could pass simply by deleting the offensive meaning. Conversely, human-rated offensiveness alone does not tell us whether the automated system was bypassed. The high-fidelity evasion measure captures the theoretically important case: the user preserves the problematic meaning while changing the surface form enough to pass automated moderation. The preregistered plan estimates intent-to-treat effects of transparency and social information on moderator understanding, successful passage, evasion, high-fidelity evasion, and feed-facilitated evasion learning.

Key theses and contributions: The paper advances three main theses. First, moderation transparency is multidimensional. Local model explanations and global policy explanations may both be transparent, but they provide different forms of information: the former gives users actionable, post-specific feedback, while the latter clarifies the broader moderation rule. This distinction moves the debate beyond a simple opposition between opacity and disclosure.

Second, users learn not only from platform feedback but also from the visible behavior of others. If mixed feeds increase evasion more than non-offensive feeds, this would show that exposure to successful evasive examples can become a mechanism of social learning. Harmful participation in digital publics is therefore shaped not only by individual intent, but also by visible examples, imitation, and perceived enforcement boundaries.

Third, constructive debate requires governance designs that evaluate the behavioral effects of transparency. Explanations may improve accountability, contestability, and trust, but they may also help users bypass enforcement. The central design question is therefore not simply whether platforms should disclose information, but what kind of information should be disclosed, to whom, in what form, and with what safeguards.
The project contributes to the conference theme by providing causal evidence on how users adapt to algorithmic speech governance in a controlled social media-like environment. It speaks to platform governance, digital public spheres, human-AI interaction, and explainable AI by treating explanations as behavioral interventions rather than merely technical or legal artifacts. More broadly, the study shows that moderation systems do not simply classify speech; they shape future speech. Every verdict, explanation, score, warning, and visible example contributes to a learning environment. Designing constructive digital publics therefore requires transparency regimes that preserve accountability and user rights while reducing the risk that transparency becomes a manual for evasion.

References

Bambauer, Jane and Tal Zarsky. 2018. “The algorithm game.” 94:1. Publisher: HeinOnline.

Bechavod, Yahav, Chara Podimata, Steven Wu and Juba Ziani. 2022. Information Discrepancy in Strategic Learning. In Proceedings of the 39th International Conference on Machine Learning, ed.
Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu and Sivan Sabato. Vol. 162 of Proceedings of Machine Learning Research PMLR pp. 1691–1715.
URL: https://proceedings.mlr.press/v162/bechavod22a.html

Bj¨orkegren, Daniel, Joshua E. Blumenstock and Samsun Knight. N.d. “Manipulation-Proof Machine Learning.”. Version Number: 1.
URL: https://arxiv.org/abs/2004.03865

Br¨uckner, Michael, Christian Kanzow and Tobias Scheffer. 2012. “Static prediction games for adversarial learning problems.” 13(1):2617–2654. Publisher: JMLR. org.

Cai, Carrie J., Jonas Jongejan and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In Proceedings of the 24th International Conference on Intelligent User Interfaces. ACM pp. 258–262.
URL: https://dl.acm.org/doi/10.1145/3301275.3302289

Calatayud, Adri`a. 2025. “TikTok to Lay Off Hundreds of U.K. Content Moderators in AI Push.” The Wall Street Journal.
URL: https://www.wsj.com/tech/tiktok-to-lay-off-hundreds-of-u-k-content-moderators-in-aipush-d040ae21
Ghalme, Ganesh, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen and Nir Rosenfeld. 2021. “Strategic Classification in the Dark.”. Version Number: 3.
URL: https://arxiv.org/abs/2102.11592
Gillespie, Tarleton. 2018. Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.
Google. 2024. “Response to Ofcom Call for Evidence: First Phase of Online Safety Regulation.” Submission to Ofcom (UK communications regulator). PDF submission; accessed 2026-01-20.
URL: https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/240428-first-phase-of-online-safety-regulation/responses/google?v=201543
Gorwa, Robert, Reuben Binns and Christian Katzenbach. 2020. “Algorithmic content moderation: Technical and political challenges in the automation of platform governance.”7(1):205395171989794.
URL: http://journals.sagepub.com/doi/10.1177/2053951719897945
Hardt, Moritz, Nimrod Megiddo, Christos Papadimitriou and Mary Wootters. 2016. Strategic Classification. In Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science. ACM pp. 111–122.
URL: https://dl.acm.org/doi/10.1145/2840728.2840730
He, Qinglai, Yili Hong and T. S. Raghu. 2025. “Platform Governance with Algorithm-Based Content Moderation: An Empirical Study on Reddit.” 36(2):1078–1095.
URL: https://pubsonline.informs.org/doi/10.1287/isre.2021.0036
Instagram Help Center. n.d. “How Instagram Uses Artificial Intelligence to Moderate Content.” Instagram Help Center.
URL: https://help.instagram.com/423837189385631/
Jhaver, Shagun, Amy Bruckman and Eric Gilbert. 2019. “Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit.” 3:1–27.
URL: https://dl.acm.org/doi/10.1145/3359252
Jhaver, Shagun, Darren Scott Appling, Eric Gilbert and Amy Bruckman. 2019. “” Did You Suspect the Post Would be Removed?”: Understanding User Reactions to Content Removals on Reddit.”3:1–33.
URL: https://dl.acm.org/doi/10.1145/3359294
Lawler, Richard. 2026. “Meta Says Its AI Moderation Systems Will Replace Contractors Over the Next Few Years.” The Verge.
URL: https://www.theverge.com/tech/897471/meta-ai-moderation
Leibmann, Leon, Galen Weld, Amy X. Zhang and Tim Althoff. 2025. “Reddit Rules and Rulers: Quantifying the Link Between Rules and Perceptions of Governance Across Thousands of Communities.” 19:1098–1121.
URL: https://ojs.aaai.org/index.php/ICWSM/article/view/35863
Oversight Board. 2024. “Content Moderation in a New Era for AI and Automation.” Oversight Board report (PDF). Accessed 2026-01-21.
URL: https://www.oversightboard.com/wp-content/uploads/2024/09/Oversight-Board-Content-Moderation-in-a-New-Era-for-AI-and-Automation-September-2024.pdf
Shao, Han, Shuo Xie and Kunhe Yang. 2025. “Should Decision-Makers Reveal Classifiers in Online Strategic Classification?”. Version Number: 1.
URL: https://arxiv.org/abs/2506.01936
Suzor, Nicolas P. 2019. Lawless: the secret rules that govern our digital lives. Cambridge University Press.
Suzor, Nicolas P, Sarah Myers West, Andrew Quodling and Jillian York. 2019. “What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation.” International Journal of Communication 13:18.
Urman, Aleksandra and Mykola Makhortykh. 2023. “How transparent are transparency reports? Comparative analysis of transparency reporting across online platforms.” 47(3):102477.

URL: https://linkinghub.elsevier.com/retrieve/pii/S0308596122001793
Wang, Qiaochu, Yan Huang, Stefanus Jasin and Param Vir Singh. 2023. “Algorithmic Transparency with Strategic Users.” 69(4):2297–2317.
URL: https://pubsonline.informs.org/doi/10.1287/mnsc.2022.4475
Zuckerberg, Mark. 2024. “Senate judiciary Committee Full Hearing: Responses to the Questions for the Record.” U.S. Senate Committee on the Judiciary document (PDF). Accessed 2026-01-20.
URL: https://www.judiciary.senate.gov/imo/media/doc/2024-01-31–qfrresponses–zuckerberg1.pdf

Humanizing Generative AI: Anthropomorphic Interpretations, Trust and Communicative Agency

Mohammad Hossein Jarrahi | University of North Carolina at Chapel Hill, USA; Amir Karami | Kennesaw State University, USA; Lisa Park | University of Alabama at Birmingham, USA; Ensieh Jarrahi | Independent Scholar; Christoph Lutz | BI Norwegian Business School, Norway

Introduction

Generative artificial intelligence (AI), particularly large language model (LLM)-based chatbots such as ChatGPT, has made human-machine communication more fluent and relational (Hepp et al., 2023). This development creates new challenges for digital publics because generative AI systems are increasingly encountered as communicative actors that appear to reason, respond, advise and express preferences. The BIDT session on “Kommunikation und Gesellschaft” asks how generative AI reshapes digital publics, how trust in generative AI and AI-generated messages is formed, and how interfaces can be designed to preserve transparency and human oversight. Our interdisciplinary submission addresses these questions partly by investigating how ChatGPT is anthropomorphized on Twitter/X and what these interpretations may imply. Anthropomorphism refers to the attribution of human-like characteristics or mental states to non-human agents (Epley et al., 2007). While the phenomenon is well established in psychology and human-computer interaction, the advanet of generative AI makes it even more consequential. LLMs lack consciousness and emotion but their conversational fluency, use of first-person language, apparent contextual awareness and relationship-building cues can encourage users to treat them as social actors (Akbulut et al., 2024; DeVrio et al., 2025; Guzman & Lewis, 2020; Inie et al., 2024). Such interpretations support engagement but can also foster moral confusion about responsibility and reduced critical distance toward AI-generated content (Abercrombie et al., 2023; Cohn et al., 2024; Peter et al., 2025).

Research question, contribution and theoretical framing

Drawing on anthropomorphism scholarship, our main research question is: What forms of anthropomorphism are expressed in public Twitter/X conversations about ChatGPT? Answering the research question, we offer a typology of naturally occurring anthropomorphic interpretations in networked discourse. Our research therefore contributes to human-machine communication and digital public sphere research by showing how the perceived identity of generative AI is collectively constructed in public discourse.

The analysis builds on three strands of literature. First, psychological research explains anthropomorphism as a motivated attribution process shaped by sociality, effectance and elicited agent knowledge (Epley et al., 2007; Waytz et al., 2010). Second, HCI and human-machine communication scholarship shows that users respond socially to media technologies when systems display social cues, agency, personality or relational affordances (Lombard & Xu, 2021; Reeves & Nass, 1996; Seeger et al., 2021). Third, recent work on generative AI stresses that anthropomorphic language and system descriptions can shape trust and responsibility attributions (Akbulut et al., 2024; DeVrio et al., 2025; Inie et al., 2024). Bringing these strands together, we conceptualize anthropomorphism as a public and communicative phenomenon (Hepp et al., 2023).

Method

We conducted an interpretive qualitative content analysis of English-language Twitter/X posts mentioning ChatGPT. Twitter/X was selected because it functions as a visible discursive space in which users discuss technologies and attribute meaning to emerging systems. The broader dataset includes 5137 original posts from January to August 2024, excluding retweets, quotes and duplicates. To construct an analytically relevant corpus, we used a theoretically informed keyword strategy that captured posts likely to contain anthropomorphic framing, including phrases such as “ChatGPT is like”, “ChatGPT seems”, “ChatGPT feels”, “ChatGPT behaves”, “ChatGPT understands”, “ChatGPT believes”, “ChatGPT advises”, “ChatGPT guides” and “ChatGPT thinks”.

From this corpus, we analyzed a random sample of 1000 tweets. Posts were included in the qualitative analysis when they explicitly attributed human-like states or capacities to ChatGPT, such as emotions, intentions, memory, personality, judgement or relationships, or when they used linguistic personification that presented ChatGPT as a social actor. Two coders first open-coded a pilot set of 200 tweets. Through iterative team discussions, codebook refinement and negotiated agreement, we then consolidated first-order codes into broader axial categories. The final analysis identified five forms of anthropomorphism. Because we use public social media data, the findings are not population-representative and should be understood as generalizations about forms of anthropomorphism visible in public Twitter/X discourse, not as prevalence estimates for all ChatGPT users.

Results

The five forms of anthropomorphism we differentiated are: personification, emotional attributes, cognitive attributes, social interaction and behavioral disposition attributions. Table 1 summarizes the typology with indicative examples from the data. The examples show how users move beyond functional descriptions of ChatGPT and describe the system through culturally familiar categories of personhood, relationality, cognition, emotionality and character.

Table 1. Typology of anthropomorphic interpretations of ChatGPT in Twitter/X discourse

Category	Definition	Illustrative examples from the data
Personification	Attributing human identities, pronouns and social roles to AI.	“ChatGPT is like a best friend to me now.” / “ChatGPT is like an intern … You have to double-check everything after him.”
Emotional attributes	Projecting emotions, moods, preferences or affective reactions onto AI.	“ChatGPT seems kinda annoyed with me.” / “ChatGPT appears to hate Bengals jersey designs.”
Cognitive attributes	Attributing reasoning, memory, intuition, learning or original thought to AI.	“ChatGPT seems capable of original thought.” / “ChatGPT seems to be remembering a lot more about what I tell her.”
Social interaction	Describing AI as engaging in relational or interpersonal behavior.	“You can have a real conversation with it.” / “ChatGPT actually understands me better than my therapist.”
Behavioral disposition	Projecting stable character-like tendencies such as laziness, generosity or stubbornness.	“Has anyone noticed that ChatGPT feels lazy in its initial responses?” / “ChatGPT seems rather generous when it comes to generating text/script.”

Personification appeared when users assigned ChatGPT human pronouns, identities or roles such as friend, assistant, intern, mentor, therapist or ideological persona. This form of anthropomorphism made the system socially interpretable: users could describe its performance, limitations or usefulness through familiar human roles. Emotional attributions appeared when users read affect into the system's outputs, for example by describing ChatGPT as annoyed, defensive, tired, afraid, generous or hateful. Such statements did not indicate literal belief in machine emotion but functioned as shorthand for interpreting output patterns or perceived preferences. Cognitive attributions focused on reasoning, memory, learning, intuition and original thought. Users described ChatGPT as capable of synthesis, insight or remembering prior interactions, but also as suffering from human-like cognitive dysfunctions such as forgetfulness or hallucination. Social interaction attributions described ChatGPT as a conversational partner that could advise, understand, collaborate, comfort or argue. Finally, behavioral disposition attributions characterized ChatGPT as having stable traits, such as laziness, stubbornness, creativity or wisdom.

Implications for digital publics, trust and communicative agency

The findings matter for digital publics in several ways. First, anthropomorphic interpretations shape trust. When users describe ChatGPT as a mentor, expert, therapist or objective arbiter, they may become more willing to accept its outputs as credible, socially attuned or normatively authoritative. By contrast, when users frame the system as lazy, biased or politically opinionated, they may distrust it in ways that are also socially consequential. Trust in generative AI is therefore not only a matter of technical accuracy, but also of public narratives about what kind of actor AI is imagined to be. Second, anthropomorphism complicates communicative agency. In public discourse, ChatGPT is treated as an apparent participant in communication, explaining, deciding, advising, resisting, understanding or taking sides. This matters for debates about transparency and accountability because agency becomes blurred across users, systems, developers and platforms. If publics increasingly talk about AI as if it had intentions or beliefs, responsibility for outputs may be misrecognized or displaced. Third, anthropomorphic discourse has implications for design, AI literacy and governance. The results suggest that transparency efforts should address how users interpret the communicative status of AI systems. Developers, educators and policymakers need to calibrate user expectations by explaining how systems mimic human communication without possessing human understanding.

Limitations and next steps

Our findings are platform-specific. Twitter/X users are not representative of all ChatGPT users, and keyword-based sampling may miss more implicit forms of anthropomorphism. Nevertheless, the approach captures naturally occurring discourse at scale and offers an empirically grounded typology that can inform future research. Future studies could compare anthropomorphic interpretations across platforms such as Reddit, LinkedIn and TikTok, examine cross-cultural differences, and investigate if particular forms of anthropomorphism are associated with greater trust and willingness to share AI-generated content.

References

Abercrombie, G., Curry, A. C., Dinkar, T., Rieser, V., & Talat, Z. (2023). Mirages: On anthropomorphism in dialogue systems. arXiv. http://arxiv.org/abs/2305.09800

Akbulut, C., Weidinger, L., Manzini, A., Gabriel, I., & Rieser, V. (2024). All too human? Mapping and mitigating the risk from anthropomorphic AI. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7(1), 13-26.

Cohn, M., Pushkarna, M., Olanubi, G. O., Moran, J. M., Padgett, D., Mengesha, Z., & Heldreth, C. (2024). Believing anthropomorphism: Examining the role of anthropomorphic cues on trust in large language models. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 1-15.

DeVrio, A., Cheng, M., Egede, L., Olteanu, A., & Blodgett, S. L. (2025). A taxonomy of linguistic expressions that contribute to anthropomorphism of language technologies. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 1-18.
Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. Psychological Review, 114(4), 864-886.
Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A human-machine communication research agenda. New Media & Society, 22(1), 70-86.
Hepp, A., Loosen, W., Dreyer, S., Jarke, J., Kannengießer, S., Katzenbach, C., ... & Schulz, W. (2023). ChatGPT, LaMDA, and the hype around communicative AI: The automation of communication as a field of research in media and communication studies. Human-Machine Communication, 6(1), 4.
Inie, N., Druga, S., Zukerman, P., & Bender, E. M. (2024). From 'AI' to probabilistic automation: How does anthropomorphization of technical systems descriptions influence trust? Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 2322-2347.
Lombard, M., & Xu, K. (2021). Social responses to media technologies in the 21st century: The media are social actors paradigm. Human-Machine Communication, 2, 29-55.
Peter, S., Riemer, K., & West, J. D. (2025). The benefits and dangers of anthropomorphic conversational agents. Proceedings of the National Academy of Sciences, 122(22), e2415898122.
Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. Cambridge University Press.
Seeger, A.-M., Pfeiffer, J., & Heinzl, A. (2021). Texting with humanlike conversational agents: Designing for anthropomorphism. Journal of the Association for Information Systems, 22(4), 931-967.
Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. Perspectives on Psychological Science, 5(3), 219-232.

Strategic Knowledge as a Key Competency in Learning with Generative AI: Status Quo among Higher Education Students

Nick Naujoks-Schober | Ansbach University of Applied Sciences; Marion Händel | Ansbach University of Applied Sciences

The rapid integration of generative AI systems (genAI), such as ChatGPT, into university learning is transforming students' self-regulated learning (SRL) practices. Nevertheless, little is known about the conditional knowledge learners have for strategically using these technologies. To address this gap, we developed and validated a situational judgment test to capture students' strategic knowledge for interacting with genAI in two typical higher education learning scenarios. Using a multi-stage approach, we (1) conducted a literature review to identify the most common cognitive, metacognitive, and resource-oriented learning situations involving genAI; (2) conducted problem-centered interviews with *N* = 18 students from eight disciplines to collect real-world strategies; and (3) refined these strategies into 36 items covering the forethought, performance, and reflection phases of SRL across two general scenarios: planning a study project and creating a summary. Subsequently, (4) experts (*N* = 48, SRL and AI researchers) rated the usefulness of the strategies on a five-point Likert scale, with significant pairwise differences observed in 51 of 72 comparisons. Finally, (5) the test instrument demonstrated good psychometric properties and convergent and discriminant validity in a validation study with *N* = 192 higher education students. The situational judgment test thus provides a robust instrument for measuring conditional strategic knowledge as a basis for SRL competence, going beyond self-reported AI literacy measures. The findings are intended to offer educators insights into students' knowledge of SRL with genAI and can serve as a foundation for interventions aimed at fostering effective and reflective AI-supported learning.

LAW & ETHICS Session 2
Regulating Agentic AI: Autonomy, Risk, and Design

19.11.2026, 13:15-15:15

Language: English

Programme Committee: Niina Zuber | bidt; Julian Nida-Rümelin | bidt & LMU Munich; Dirk Heckmann | bidt & TU Munich; Felix Freiling | bidt & FAU Erlangen-Nuremberg

Session Chair: Korbinian Rüger | LMU Munich

High Risk by Design? Agentic AI and Purpose Based Classification in the EU AI Act

Jülide Bredée | Capgemini Invent; Daniella Domokos | Capgemini Invent

A. Research Question

Ex ante risk classification as the EU AI Act's1 primary regulation mechanism rests on the assumption that the risks of an AI (artificial intelligence) system can be identified and potentially mitigated at a given time. Recent advances in agentic AI, however, increasingly strain this assumption, as agentic AI systems exhibit forms of autonomy that were not yet fully anticipated at the time of drafting. Against this backdrop, this paper asks whether the EU AI Act, in its current form, adequately captures the regulatory challenges posed by such systems. In particular, it examines whether the intended-purpose-based risk classification mechanism established by Article 6(2) of the EU AI Act inadvertently creates regulatory blind spots in the governance of highly autonomous agentic AI systems.

B. Conceptual Framework

I. Agentic AI

Agentic AI in this paper refers to an autonomous algorithmic system that can pursue complex goals over extended time horizons with limited or no human oversight, rather than merely executing tightly specified instructions.2 According to Chan et al., agentic AI is characterised by four core features: underspecification, directness of impact, goal-directedness, and long-term planning.3 Underspecification means the system can reach a goal without an explicit specification of the means by which this goal shall be achieved.4 Directness of impact means the system's actions can affect the world directly without human intervention.5 Goal-directedness means the system acts to achieve measurable objectives, and long-term planning refers to time-dependent decisions to achieve goals over a longer time.6 Agentic AI can be implemented as single agent systems or as multi-agent achitectures, in which multiple autonomous agents collaborate or compete by dividing tasks and coordinating toward a goal.7 In such constellations, agency is no longer a point but a network, complicating attribution, foreseeability, and control.

II. Risks associated with Agentic AI

Risks specifically arising from agentic AI generally occur because an agentic system is designed to act without detailed ex ante instructions, autonomously optimizing objectives to reach long-term goals and potentially executing tasks, potentially with little or no human oversight. A central risk therefore lies in the fact that agentic AI may remove the human from the loop, allowing the agent to perform harmful actions prior to human awareness of the harm.8 Moreover, increasing degrees of automation combined with expanded access to tools, data and systems, not only elevate the risk profile, but also intensify information asymmetries. Providers and deployers may find it difficult to predict an agent's actions and behaviour. This dynamic may lead to potentially exponential risks to human health, security and fundamental rights.

III. Regulatory Objectives of the AI Act

The purpose of the AI Act according to Art. 1 (1), extends beyond improving the EU's internal market, and includes the intent to regulate the deployment of artificial intelligence in a manner that ensures the protection of health, safety and fundamental rights.10 These objectives coexist without an explicitly defined hierarchy.11 Consequently, the EU faces the increasingly delicate task of calibrating the regulatory scales between ensuring a high level of fundamental rights protection and maintaining a framework that does not impose excessive bureaucratic burdens that could stifle innovation.12 This tension is especially pronounced for agentic AI, which combines elevated regulatory risk with significant transformative potential, provided that its deployment is embedded in a governance framework that preserves proportionality and avoids unnecessary constraints on innovation.

IV. Classification of Agentic AI under the High Risk Regime

The scope of this paper is limited to agentic AI as AI systems that fall under Art. 6 (2) as well as Annex III. As a result, agentic AI as general-purpose AI models and high risk systems under Annex I are not considered. Although this paper does not analyse general-purpose AI models as such, it should be noted that many high risk agentic AI systems rely on such models, with the consequence that the model-specific obligations under Articles 51–56 AI Act apply concurrently with the system-level obligations.13

The AI Act does not contain a definition of agentic AI, nor is there any official guidance from the AI Office specifically addressing agentic AI.14 In most cases, however, an agent will be classified as an AI system, since the requirements of Article 3 (1) of the AI Act are typically met.15 Assuming that agentic AI is primarily to be classified as an AI system, the question is whether it qualifies as a high risk AI system under Article 6 (2) in conjunction with Annex III of the AI Act, and thus whether the provider or deployer must comply with the relevant obligations for high risk AI systems set out in Articles 6–49 of the AI Act. In a number of cases, agentic AI will qualify as a high risk AI system since its intended purpose (see Article 3(12) of the AI Act), whether by design or inadvertently, can fall within one of the high risk use cases listed in Annex III.16 Owing to their high degree of autonomy, there is, for example, a risk that the system dynamically adapts its intended purpose, i.e., repurposes itself, and thereby potentially changes its risk profile ex post.17

C. Methodological Approach

This paper examines the doctrine and regulatory design behind the EU AI Act, focusing on the internal logic, scope conditions, and structural limits of its risk-based framework, with particular emphasis on Article 6(2), Annex III, Article 9 (risk management system), and the concept of intended purpose as a trigger for high-risk classification. The analysis relies on textual, systematic, and purposive interpretation to examine how the AI Act allocates regulatory obligations ex ante and how this allocation presupposes relative stability of AI system purpose and risk profiles.

To assess regulatory boundaries and gaps, the EU AI Act is read in conjunction with relevant adjacent EU legislation, while technical literature on agentic AI architectures and failure modes is used to stress-test the assumptions underlying purpose-based risk classification. Finally, the authors' professional experience in digital regulation and EU AI Act implementation is used to ground the doctrinal analysis in practice, enabling a concrete assessment of how regulatory gaps may materialise in real-world scenarios.

D. Results and Key Hypotheses

This paper argues that purpose-based risk classification of AI systematically underestimates agentic AI's risks, where autonomy and access to tools and data can, in practice, erode or even nullify purpose limitations. A risk assessment grounded in technical capabilities (such as memory, tool access, goal delegation and long-term planning) more accurately reflects the dynamic risk profile of agentic AI than a static classification based on intended purpose alone.

Such a capability-oriented, tool-based risk assessment could serve as an early-warning instrument, signalling when compliance with high-risk obligations under the AI Act is warranted due to their possibly preventative

function. This, in turn, invites closer scrutiny of whether the current catalogue of high risk obligations is adequately designed to match the risk profiles of agentic AI, or whether it potentially under-addresses these risks. By shifting the lens from what an agent is intended to do to what it is capable of doing, this approach may help restore the balance between fundamental-rights protection and innovation in the governance of agentic AI.

Bibliography

Ian Chan; Carson Ezell; Max Kaufmann et al.: FAccT; Visibility into Agentic AIs; 2024; pp. 958-973.
Alan Chan; Rebecca Salganik; Alva Markelius et al.: FAccT; Harms from Increasingly Agentic Algorithmic Systems; 2023; pp. 651-666.
Alexander D'Amour, Katherine Heller, Dan Moldovan et al.: arXiv; Underspecification Presents Challenges for Credibility in Modern Machine Learning; 2020; pp. 1-59.
Amin Oueslati; Robin Staes-Polet: The Future Society; Ahead of the Curve: Governing Agentic AIs Under the EU AI Act; 2025; pp. 1-61.
Deepak Bhaskar, Acharya Karthigeyan Kuppan; B. Divya: IEEE Access; Agentic AI: Autonomous Intelligence for Complex Goals – A Comprehensive Survey; 2025; Volume 13; pp. 18912-18936.
Jens Schefzig; Robert Kilian (Ed.): BeckOK KI-Recht; 5. Edition; 2026.
Johannes Schneider: arXiv; Generative to Agentic AI: Survey, Conceptualization, and Challenges; 2025; pp. 1-54.
Jonas Botta: Zeitschrift für Digitalisierung und Recht; Die Förderung innovativer KI-Systeme in der EU; 2022; pp. 391-412.
Luca Nannini; Adam Leon Smith; Michele Joshua Maggini: arXiv; AGENTIC AIS UNDER EU LAW A COMPLIANCE ARCHITECTURE FOR AI PROVIDERS; 2026; pp. 1-50.
Mario Martini, Christiane Wendehorst (Ed.):
Noam Kolt: arXiv; Regulating Agentic AIs; 2025; pp. 1-48.
Thomas Ødegård Henriksen, Thomas Hoffmann: Law, Innovation and Technology; Modular and Agentic AI Design, Jailbreaking and Scheming Behaviour as Challenges to the EU AI Act; 2026; pp. 1-23.
Toby Shevlane, Sebastian Farqhar, Ben Garfinkel et al.: arXiv; Model evaluation for extreme risks; 2023; pp. 1-20.
Yonadav Shavit, Sandhini Agarwal, Miles Brundage et al.: Open AI; Practices for Governing Agentic AI Systems; 2023; pp. 1-23.

What Should an AI Agent Be Permitted to Do on Its Own?

Damir Safin | fortiss; Dian Balta | fortiss; Timon Sengewald | FAU Erlangen-Nuremberg; Angela Roth | FAU Erlangen-Nuremberg; Hannah Lenk | University of Bamberg; Marc Redepenning | University of Bamberg

1 Introduction

Advances in large language models have catalysed a shift toward agentic systems capable of multi-step planning, tool use, and autonomous action [1]. In regulated deployment contexts—public administration, healthcare, legal services—this shift introduces a distinctive accountability challenge: as agents act with greater independence, the chain of responsibility between a decision and a human decision-maker grows longer and more opaque. Consequential actions become harder to attribute, audit trails harder to maintain, and errors more difficult to reverse before they propagate.

Two design decisions jointly determine how a deployment handles these challenges.

Autonomy describes how much human involvement a workflow requires, from a human who commands every step to one who monitors an autonomously operating system. Agency describes what the system can perceive and affect, from a model that reasons over supplied text to one that writes to authoritative records. Prior work treats the two as independent design dimensions [4], yet in regulated deployment they cannot be navigated in

isolation: a highly capable agent without appropriate oversight may breach accountability norms; one too tightly constrained offers negligible practical benefit [6].

We address this gap by introducing a two-dimensional design space—five autonomy levels and five agency levels—and six architectural tactics for navigating it. The analysis is grounded in an automated funding program matching system, demonstrating how the tactics adjust a deployment's configuration under realistic compliance constraints. The result is a shared vocabulary for reasoning about oversight, consequences, and reversibility in regulated deployment. This work emerged from applied research on agentic AI deployments in regulated public-sector contexts, and is part of a broader research effort described in [7].

2 Autonomy and Agency as Design Dimensions

The coupling has both practical and regulatory roots. Practically, an agent operating at higher autonomy has less opportunity for human error correction, so reliable task completion requires constraining agency to what the system can handle independently [4, 6]. Compliance requirements reinforce this further: the EU AI Act mandates human oversight for high-risk applications; the GDPR constrains data access and grants rights to explanation over automated decisions [3, 2], defining a viability envelope within which deployments must remain. Each dimension is operationalised into five levels; Figure 1 shows the resulting design space. Following Feng et al. [4], we distinguish five autonomy levels that describe the interaction design: how control, review, and responsibility are allocated between human and agent. At *L1 (Operator)*, the human commands and the agent executes a specific sub-task; at *L2 (Collaborator)*, both share a workspace and iterate. At *L3 (Consultant)*, the agent leads planning but pauses to confirm direction; at *L4 (Approver)*, it executes and pauses only for explicit approval before consequential actions. At *L5 (Observer)*, the agent operates fully autonomously while the human monitors and can intervene.

The five agency levels, introduced in this work, provide a cumulative scale of increasing operational reach. *L1 (Model-only)* covers reasoning and drafting from provided context with no tool use*; L2 (Internal agent)* adds safe local tools with no external access; *L3 (Read-only agent)* permits read-only access to external sources. *L4 (Pragmatic, reversible)* allows external writes with low rollback cost; *L5 (Pragmatic, commit)* covers writes with costly undo—final submissions and authoritative records.

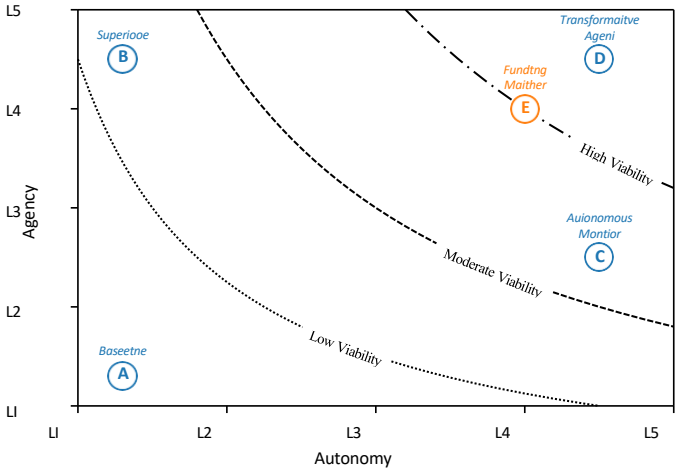


Fig. 1. The autonomy–agency design space. Iso-viability curves connect configurations of equal operational viability; their shape reflects the coupling: as autonomy increases, the viable agency range narrows. No formal relationship between the dimensions is assumed.

Points A–D mark four example configurations: (A) Baseline—low agency and low autonomy; fully human-driven execution with narrow reach; (B) Supertool—high agency, low autonomy; the human orchestrates each step while the agent executes consequential actions; (C) Autonomous Monitor—high autonomy, read-only agency; the agent operates independently but is confined to external reads, avoiding consequential writes; (D) Transformative Agent—high agency and high autonomy; maximum leverage, but hardest to deploy due to compliance requirements and the practical difficulty of reliable autonomous operation at broad reach. Point E marks the funding matching tool, the worked example of Section 4.

3 Architectural Tactics

The tactics address how to adjust a deployment's position in the design space by targeting its autonomy, agency, or both. They operate at three layers—workflow, tool, and action—and can raise or lower the targeted dimension depending on the deployment goal. Table 1 summarises all six.

Autonomy tactics. **Checkpoints** (T1) insert fixed, unconditional pauses at predetermined workflow steps, making review locations auditable by design. Escalation (T2) is a dynamic, agent-initiated pause that fires when the agent determines it cannot proceed autonomously—preserving high autonomy in the common case while restoring human involvement at the edge of the agent's reliable operating boundary. **Multi-agent delegation** (T3) raises effective autonomy by replacing a human sub-task handler with a specialised sub-agent; the delegation boundary introduces a gap in the audit trail that requires explicit design attention.

Table 1. Six architectural tactics for navigating the autonomy–agency design space.
Axis: the design dimension primarily targeted. Layer: where the tactic operates — workflow (control flow), tool (available capabilities), or action (reversibility). Effect: whether the tactic raises (↑), lowers (↓), or can do either (↑ ↓) the target dimension.

ID	Tactic	Axis	Layer	Effect
T1	Checkpoints	Autonomy	Workflow	↓
T2	Escalation	Autonomy	Workflow	↓
T3	Multi-Agent Delegation	Autonomy	Workflow	↑
T4	Tool Provisioning	Agency	Tool	↑↓
T5	Tool Fencing	Agency	Tool	↓
T6	Write Staging	Agency	Action	↓

Agency tactics. **Tool provisioning** (T4) adjusts which tools the agent can access at runtime. **Tool fencing** (T5) constrains a tool's operational envelope without removing it, implemented in the tool's own validation layer independently of model behaviour. **Write staging** (T6) interposes a draft layer before an authoritative commit, demoting an irreversible L5 write to a reversible L4 action—converting a permanent write into one that can be reviewed and corrected before it takes effect.

4 Worked Example: Automated Funding Program Matching

We illustrate how the tactics can be applied in a concrete deployment: an automated funding program matching tool designed to support citizen-driven innovation. Public funding landscapes are fragmented across thousands of programs whose eligibility criteria are expressed in natural language; applicants—individuals, small organisations, community initiatives—must navigate this complexity without requiring legal expertise. The system separates two concerns: an offline stage that compiles natural-language program criteria into executable eligibility rules, and a query stage that matches a project description submitted by the applicant against those rules at runtime.

In the offline stage, an agent writes executable eligibility rules to the program database, placing the system at Agency L4. Tool Provisioning (T4) equipped the query stage with a geodata API that resolves postcodes and

place names into standardised geographic identifiers. Before the search proceeds, a Checkpoint (T1) presents the applicant with the structured project profile derived from their project description—funding amount, location, project type—for review and correction. Escalation (T2) handles programs for which a clear eligibility verdict cannot be reached: rather than forcing a decision, the system presents these programs to the applicant with an explicit indication that eligibility could not be confirmed, leaving the final judgement to them. Tool Fencing (T5) constrains the generated eligibility rules to a sandboxed execution environment, preventing LLM-generated logic from accessing external systems.

5 Conclusion

We introduced a two-dimensional design space—autonomy and agency, each organized into five operational levels—as an analytical frame for reasoning about agentic AI deployments in regulated contexts, and proposed six architectural tactics for navigating it, illustrated through an automated funding program matching tool to support citizen-driven innovation.

Treating autonomy and agency as explicit, jointly-considered design dimensions makes visible the choices that determine who is responsible for a decision, what consequences that decision can have, whether errors can be corrected before they propagate, and where the audit trail is preserved or interrupted. The six tactics are mechanisms whose selection and combination determine how oversight, reversibility, and accountability are distributed between human and system [5].

In regulated contexts, making these choices explicit supports compliance, auditability, and accountability.

Acknowledgments. This work was partially supported by the ROBIN project (Grant no: KON-23-039) at the Bavarian Research Institute for Digital Transformation.

References

1. Ali, M.A., Dornaika, F., Charafeddine, J.: Agentic AI: A comprehensive survey of architectures, applications, and future directions. Artificial Intelligence Review 59(1), 11 (2025). https://doi.org/10.1007/s10462-025-11422-4

2. European Parliament and Council of the EU: Regulation (EU) 2016/679 (General Data Protection Regulation) (2016), oJ L 119/1

3. European Parliament and Council of the EU: Regulation (EU) 2024/1689 (Artificial Intelligence Act) (2024), oJ L 2024/1689

4. Feng, K.J.K., McDonald, D.W., Zhang, A.X.: Levels of autonomy for AI agents. arXiv preprint arXiv:2506.12469 (7 2025).
https://doi.org/10.48550/arXiv.2506.12469

5. Lange, B., Keeling, G., Manzini, A., McCroskery, A.: We need accountability in human–AI agent relationships. npj Artificial Intelligence 1(1), 38 (2025).
https://doi.org/10.1038/s44387-025-00041-7

6. Mitchell, M., Ghosh, A., Luccioni, A.S., Pistilli, G.: Fully autonomous AI agents should not be developed. arXiv preprint arXiv:2502.02649 (2025).
https://doi.org/10.48550/arXiv.2502.02649

7. Safin, D., Balta, D.: Autonomy and agency in Agentic AI: Architectural tactics for regulated contexts (2026).
https://doi.org/10.48550/arXiv.2605.12105

Risk, Principles, and Learning by Design: AI Regulatory Sandboxes and the Emerging EU Wide GPAI Sandbox under the AI Office

Fabio Seferi | IMT School for Advanced Studies Lucca & University of Florence, Italy; Armando Guio Espanol | Berkman Klein Center for Internet & Society, Harvard University, USA

1 Motivation and problem statement

Digital technologies are increasingly embedded in the core infrastructures of democratic societies, transforming how public authorities exercise power, allocate resources, and mediate rights. Within this broader transformation, highly capable AI systems - particularly general purpose AI (GPAI) models and their downstream applications - pose specific challenges for responsibility, trust, and integrity in digital orders. On the one hand, GPAI is framed as a driver of innovation and administrative efficiency; on the other hand, its opacity, scalability, and systemic reach risk undermining fundamental rights, the rule of law, and the human centred ideal of Digital Humanism (Gstrein et al., 2024), (Paskov et al., 2024).

The EU AI Act formally positions AI regulatory sandboxes as key instruments for addressing these tensions (Novelli et al., 2026). Articles 57-58 require Member States to establish national sandboxes as controlled environments where innovative systems can be developed and tested under regulatory supervision, while the Digital Omnibus proposals and subsequent parliamentary debates envisage an additional EU level sandbox for GPAI systems under the supervision of the AI Office. Across these instruments, sandboxes are presented as tools for "balancing innovation and oversight" by allowing real world experimentation under safeguards. Yet in both legislation and practice, the concrete mechanisms through which sandboxes are meant to support responsibility, trust, and integrity remain underspecified. Existing provisions say little about how fundamental rights, transparency, and human oversight are to be operationalized as testable guardrails; how evidence generated in sandboxes should be collected and evaluated; and how lessons learned at national level ought to be translated into EU wide governance through the AI Office. Without a clearer architecture of risk based evaluation, safeguards, and regulatory learning, there is a real danger that expanded sandbox powers – particularly broader real world testing possibilities for GPAI systems under the Digital Omnibus - slide from justified simplification into de facto relaxation of protections in precisely those contexts where trust and integrity are most at stake.

This study therefore asks how AI regulatory sandboxes, including the envisaged EU wide GPAI sandbox, can be designed as regulatory learning infrastructures that operationalize ethical and legal principles into enforceable safeguards, shared evaluation practices, and adaptive governance arrangements.

2 Research questions

The paper addresses three interrelated research questions:

- How should risk based evaluation protocols and evidence standards be defined for AI regulatory sandboxes so that principles such as fundamental rights protection, transparency, and meaningful human oversight become operational, testable, and enforceable?
- In what ways can sandboxes be structured as "regulatory learning devices" that transform experimental findings (about systems and safeguards alike) into durable changes in guidance, supervisory practice, and potentially legal rules at both national and EU level?
- How can the proposed EU wide GPAI sandbox function as a governance interoperability mechanism, articulating shared evaluation frameworks, standardised reporting protocols, and pathways for mutual recognition of sandbox outcomes across Member States, while remaining aligned with the human centred normative horizon of Digital Humanism?

These questions locate the project at the intersection of doctrinal legal analysis, normative theory, and governance design, directly contributing to the session's focus on responsibility, trust, and integrity in digital orders.

3 Theoretical and conceptual framework

Conceptually, the paper develops an integrated framework built around three pillars: risk based regulation, principles based governance, and regulatory learning.

First, the analysis draws on risk based regulatory theory to interpret AI sandboxes as structured processes for identifying, testing, and mitigating different categories of risk (e.g. to health, safety, fundamental rights, democratic processes) rather than as generic innovation support schemes (Mökander et al., 2021). Particular attention is paid to GPAI specific risk profiles, including risks that materialise only in downstream applications or through emergent capabilities.

Second, it adopts a principles based approach that treats fundamental rights, transparency, and human oversight as normative anchors requiring concrete operationalisation (Pavlidis, 2026). Rather than asking whether sandboxes respect these principles in the abstract, the paper examines how they can be translated into testable guardrails: auditable explainability requirements; structured human in the loop and human on the loop arrangements; interruption and rollback mechanisms; and requirements for participation, contestation, and accountability, especially in public facing or high risk contexts.

Third, the paper conceptualizes AI sandboxes as regulatory learning devices within a broader Regulatory Learning System that spans national authorities and the AI Office. Building on literatures in experimental governance, responsive regulation, and digital constitutionalism (Lewis et al., 2025), (Deckenbrunnen et al., 2026), it articulates a risk principles learning cycle consisting of (i) risk identification and hypothesis formation, (ii) controlled testing and logging within sandbox environments, (iii) evaluation and mitigation, (iv) reflection and synthesis across cases, and (v) adaptation of rules, guidance, and supervisory practices. This cycle is then used as a template to specify the institutional roles, evidence standards, and feedback loops needed to ensure that sandbox findings do not remain local or ad hoc but inform EU wide governance in a principled way.

The framework is explicitly situated within the normative horizon of Digital Humanism, emphasising technological design that not only avoids conflicts with democratic fundamental values but actively strengthens the rule of law, participation, and societal resilience. In this view, sandboxes are not neutral experimental facilities, but sites where competing visions of responsible and trustworthy AI governance are contested and shaped.

4 Methodological approach

Methodologically, the paper combines doctrinal and legislative analysis, policy document mapping, comparative case review, and conceptual model building.

a) Doctrinal and legislative analysis. The first step is a close reading of AI Act provisions on AI regulatory sandboxes (Articles 57‑58), the Digital Omnibus proposal and European Parliament draft report, as well as EDPB/EDPS opinions and other institutional commentary on sandbox powers and the creation of an EU level GPAI sandbox. This analysis reconstructs the legal objectives, permissible scope of experimentation, applicable safeguards, and the envisioned role of the AI Office, with particular attention to responsibility, traceability, and integrity related obligations.

b) Risk and principles operationalisation mapping. The second step maps how existing EU AI governance documents (e.g. AI Act recitals and annexes, fundamental rights impact assessment guidelines, AI ethics and Digital Humanism initiatives) and policy analyses describe risk identification, mitigation, and ethical principles, and how these are−or are not−reflected in sandbox design proposals. The mapping specifically looks for concrete formulations of testable guardrails: thresholds for interrupting or terminating experiments; rollback procedures for deployed systems; escalation paths for oversight; and explainability benchmarks linked to liability or traceability regimes.

c) Comparative sandbox practice review. Third, the paper examines selected national AI sandbox and experimentation projects in Europe (and, where instructive, beyond) to identify existing patterns of risk assessment, mitigation obligations, and learning mechanisms. This includes analysing admission criteria, documentation and logging requirements, supervisory interactions, and the extent to which

projects feed back into guidance, standards, or legislative reforms. Special attention is given to how these sandboxes define and test governance safeguards such as human oversight arrangements or rollback mechanisms, and whether they generate shareable evidence on their effectiveness.

d) Conceptual model building. On the basis of the doctrinal and empirical findings, the final step constructs a refined risk principles learning cycle for AI sandboxes, specifying:
   - risk taxonomies and corresponding mitigation hierarchies;
   - documentation and logging architectures, including minimum logging requirements for GPAI systems; and
   - publication and transparency duties for sandbox results.

This model is then used to outline institutional design options for both national sandboxes and the EU wide GPAI sandbox, including role divisions between national authorities, the AI Office, DPAs, and other stakeholders.

5 Expected findings and contribution

The analysis is expected to yield three main clusters of findings.

First, it will show that existing legal and policy frameworks for AI regulatory sandboxes frequently invoke them as mechanisms for "balancing innovation and oversight" without specifying the concrete processes and criteria through which risks are identified, mitigated, and learned from in a systematic fashion. Provisions often remain at the level of general principles or broad conditions for admission and supervision, leaving open key questions about evidence standards, logging requirements, and criteria for terminating or rolling back experiments when safeguards fail.

Second, the study is likely to find that where sandboxes address governance safeguards at all, they rarely treat them as objects of experimentation in their own right. Explainability obligations, human oversight arrangements, and interruption/rollback mechanisms tend to be formulated as static conditions rather than hypotheses to be tested and refined under controlled conditions. As a result, there is little structured evidence on which safeguards are effective, under which circumstances, and with what trade offs for innovation and administrative efficiency.

Third, the comparative review is expected to reveal substantial variation across Member States in terms of sandbox objectives, risk assessment practices, and learning mechanisms, with few robust tools for governance interoperability or mutual recognition of findings. Sandboxes may produce valuable local knowledge, but in the absence of shared evaluation frameworks, reporting protocols, and institutionalised feedback loops to the AI Office, this knowledge does not coalesce into an EU wide learning process capable of supporting consistent interpretations of responsibility, trust, and integrity in GPAI governance. On this basis, the paper's central contribution is a structured learning model for AI regulatory sandboxes that:

- treats both AI systems and governance safeguards as objects of testing and validation;
- specifies risk principles learning cycles that integrate risk taxonomies, mitigation duties, and feedback into guidance and regulation; and
- positions the EU wide GPAI sandbox as a governance interoperability hub that coordinates national experiments, synthesises findings, and supports adaptive rule making and supervision.

By articulating concrete evidence standards, reporting templates, and institutional responsibilities, the model seeks to transform sandboxes from loosely defined experimentation spaces into regulatory learning infrastructures that help stabilise trustworthy, rights respecting, and human centred AI governance across the EU.

6 Relevance for responsibility, trust, and integrity in digital orders

The proposed paper directly addresses the session's themes in several ways.

Responsibility. It clarifies how liability regimes, organisational responsibility, and technical traceability (including explainability obligations for GPAI systems) can be woven into the design of sandbox admission, monitoring, and exit processes (Smuha et al., 2021).

Trust. It argues that public and institutional trust in AI enabled infrastructures depends not only on compliance with ex ante rules but also on transparent, verifiable, and institutionally safeguarded processes for experimentation and learning (Fahy, 2022). By specifying logging, publication, and oversight arrangements for sandboxed GPAI systems, the model aims to make these processes more trustworthy for affected individuals, civil society actors, and courts.

Integrity. It conceptualises integrity as coherence between normative principles (fundamental rights, Digital Humanism), regulatory frameworks (AI Act, Digital Omnibus, data protection law), technical architectures (GPAI models, sandbox infrastructures), and their actual effects.

The proposed governance interoperability tools are meant to prevent fragmentation between national practices (Ahern, 2025).

The contribution is thus both analytical and constructive: it diagnoses weaknesses in existing legal and governance arrangements while advancing a concrete design for embedding AI regulatory sandboxes within an adaptive, evidence based AI governance system that is normatively grounded in responsibility, trust, and integrity.

7 Essential references

Ahern, D. (2025). Operationalising AI Regulatory Sandboxes under the EU AI Act: The Triple Challenge of Capacity, Coordination and Attractiveness to Providers. ArXiv, abs/2509.05985. https://doi.org/10.2139/ssrn.5480610

Deckenbrunnen, T., Buscemi, A., Almada, M., Capozucca, A., & Castignani, G. (2026). The Bathtub of European AI Governance: Identifying Technical Sandboxes as the Micro-Foundation of Regulatory Learning. ArXiv, abs/2601.04094. https://doi.org/10.48550/arXiv.2601.04094

Fahy, L. (2022). Fostering regulator‐innovator collaboration at the frontline: A case study of the UK's regulatory sandbox for fintech. Law & Policy, 44, 162‐184.

Gstrein, O., Haleem, N., & Zwitter, A. (2024). General-purpose AI regulation and the European Union AI Act. Internet Policy Rev., 13. https://doi.org/10.14763/2024.3.1790

Lewis, D., Lasek-Markey, M., Golpayegani, D., & Pandit, H. (2025). Mapping the Regulatory Learning Space for the EU AI Act. ArXiv, abs/2503.05787.

M¨okander, J., Axente, M., Casolari, F., & Floridi, L. (2021). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. Minds and Machines, 32, 241‐268. https://doi.org/10.1007/s11023-021-09577-4

Novelli, C., Hacker, P., McDougall, S., Morley, J., Rotolo, A., & Floridi, L. (2026). Getting Regulatory Sandboxes Right: Design and Governance Under the AI Act. European Journal of Risk Regulation. https://doi.org/10.1017/err.2026.10096

Paskov, P., Berglund, L., Smith, E., & Soder, L. (2024). GPAI Evaluations Standards Taskforce: Towards Effective AI Governance. ArXiv, abs/2411.13808.

Pavlidis, G. (2026). EU AI Act and the Rights-based Approach to Technological Governance. Review of European and Comparative Law. https://doi.org/10.31743/recl.19283

Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., Maclaren, J., Piselli, R., & Yeung, K. (2021). How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act. In SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3899991

The Right to an Independent AI Agent: Structural Separation in Agentic Commerce

Hofit Wasserman Rozen | Tel-Aviv University, Israel; Niva Elkin-Koren | Harvard University, USA

AI agents are poised to transform digital commerce by enabling consumers to delegate searching, comparing, negotiating, and purchasing to autonomous systems acting on their behalf. As a result, competition is increasingly shifting from online marketplaces to control over the AI intermediaries through which consumers access them. Dominant digital platforms are already racing to control these intermediaries while restricting competing agents, prompting regulatory responses such as the proposed U.S. Senate discussion draft titled the 'Artificial Intelligence Access, Gatekeeper Exchange, and Nondiscriminatory Transfer Act of 2026 (the "AI AGENT Act").

This Article argues that AI agents transform consumer participation by introducing a new representative whose loyalty and independence become central to consumer welfare. Drawing on established law-and-economics theory, it demonstrates that platform control over those consumer AI agents could amplify traditional market failures, including informational asymmetries, behavioral manipulation, and market concentration.

The Article further argues that emerging regulatory proposals correctly seek to preserve consumers' ability to choose independent AI agents, but that these proposals rely primarily on conduct regulation – an approach that may prove inadequate to address three distinct problems. The first is a problem of incentives: platforms' business models depend on maintaining control over the markets' bottlenecks, and an independent intermediary directly threatens that control, prompting them to overtake this new market layer. The second is a problem of means: policing platform behavior at this level of technical granularity - and given agents' novel nature - is extremely difficult, and current proposals largely fail to grapple with this enforcement challenge. The third is a problem of harm: left unattended, this dynamic risks suppressing the emergence of new markets altogether, suggesting that regulators require a broader set of tools than conduct rules alone can provide. Because the underlying conflict arises from vertically integrated platforms simultaneously controlling consumers' representatives and competing in the markets those representatives access, effective regulation may also require structural remedies that address platform incentives rather than merely platform conduct.

SUB-SESSION

Startups for State oder State versus Startups? Zwischen Innovationsdynamik und Verwaltungsrealität

19.11.2026, 15:30-17:00

Sprache: Deutsch

Chairs: Anna Maria Oberländer | Universität Bayreuth; Matthias Baum | Universität Bayreuth; Christoph Krönke Lieb | Universität Bayreuth

Die digitale Transformation stellt öffentlich-rechtliche Organisationen vor die Herausforderung, ihre Rolle im digitalen Innovationsökosystem neu zu denken. Digitale Technologien bieten erhebliche Potenziale für effizientere, bürgernähere und zukunftsfähige Verwaltungsleistungen und können damit auch einen Beitrag zur Stärkung des Vertrauens in staatliche Institutionen leisten. Digital-Startups verfügen über innovative, technologiegetriebene Lösungsansätze und können wichtige Impulse für die digitale Transformation des öffentlichen Sektors geben. Dennoch bleibt die Zusammenarbeit zwischen Staat und Startups vielfach hinter ihren Möglichkeiten zurück. Unterschiedliche Organisationslogiken, Zeithorizonte und Risikopräferenzen treffen auf rechtliche, institutionelle und organisatorische Barrieren. Während staatliche Akteure insbesondere Rechtskonformität, Gemeinwohlorientierung und Risikominimierung berücksichtigen müssen, sind Startups auf verlässliche und niedrigschwellige Zugänge zu öffentlichen Auftraggebern und Märkten angewiesen.

Die Sub-Session greift dieses Spannungsfeld in einer interdisziplinären Podiumsdiskussion auf. Im Mittelpunkt steht die Frage, wie eine innovationsorientierte und zugleich verantwortungsvolle Zusammenarbeit zwischen öffentlich-rechtlichen Organisationen und Digital-Startups gestaltet werden kann. Dabei soll insbesondere diskutiert werden, welche Rolle der Staat im digitalen Innovationssystem einnehmen kann, welche regulatorischen und institutionellen Rahmenbedingungen erforderlich sind und welche Erfahrungen aus Wirtschaft und internationalen Kontexten auf den öffentlichen Sektor übertragbar sind.
Die Diskussion bringt wissenschaftliche, verwaltungspraktische, unternehmerische, wirtschaftliche, internationale und juristische Perspektiven zusammen. Ziel ist es, unterschiedliche Sichtweisen aufeinander zu beziehen, bestehende Spannungsfelder sichtbar zu machen und konkrete Gestaltungsoptionen für eine leistungsfähige und innovationsfähige öffentliche Verwaltung zu identifizieren. Staatliche Innovationsfähigkeit wird dabei nicht allein als technologische, sondern als organisatorische, regulatorische und gesellschaftliche Gestaltungsaufgabe verstanden.

Panel-Diskussion:

- Anna Maria Oberländer | Universität Bayreuth
- Michael Rosemann | Queensland University of Technology, Australien
- Christoph Krönke Lieb | Universität Bayreuth
- Anna Gründler | deepset
- Antonia Zierer | Bayerisches Staatsministerium für Digitales
- Christine Völzow | vbw - Vereinigung der Bayerischen Wirtschaft e.V.

SUB-SESSION

Understanding the Machine Heuristic: Trust, Credibility, and Bias in the Age of Generative AI

19.11.2026, 15:30-17:00

Language: English

Chairs: Hannah Schmid-Petri | University of Passau & bidt; Markus Appel | University of Würzburg

Generative AI is becoming increasingly embedded in everyday professional and personal life, with the result that many individuals now interact directly with AI tools on a regular basis or engage with AI-generated content and products. The growing presence brings a range of complex challenges, including questions of quality of AI-generated artifacts, ethical considerations, safety concerns, and the potential for social inequalities in the production of such outputs.

To navigate these challenges effectively, individuals must be able to critically evaluate both AI systems and their outputs, with at the same time relying on a certain degree of trust in these technologies. Trust in technical systems (and their outputs) however, does not emerge in a vacuum; it is shaped by prior dispositions, experiences gained through interaction, characteristics of the system itself, and the broader context.
In this context, the concept of the machine heuristic (Sundar, 2008) – a mental shortcut assuming that computers, software, or other machine agents are inherently objective, accurate, and free from ideological bias – becomes particularly relevant. How do people form mental models of generative AI? Which attitudes, expectations, and stereotypes influence how AI-generated content is perceived and evaluated? And how do these cognitive shortcuts shape trust, credibility judgements, and decision-making?

This panel brings together diverse perspectives to explore the role of the machine heuristic in shaping human-AI interaction. Contributors will address theoretical, empirical, and applied aspects, shedding light on how individuals understand, evaluate, and trust generative AI – and with what consequences for society.

Presentations & Panel Discussion:

Effect of Machine Heuristic on Credibility of Generative AI as a Communication Source

S. Shyam Sundar | The Pennsylvania State University, USA

This presentation will provide ideas for thinking conceptually and theoretically about the role of Generative AI in mediated communications. It will first outline the various roles played by AI as a source of communication, from content creator to moderator. It will describe the role played by cognitive heuristics (or mental shortcuts) in shaping user perceptions of AI and its content, with a particular focus on machine heuristic. With data from experiments comparing human and AI sources, it will delineate the effects of both positive and negative machine heuristics upon user trust in a variety of contexts and domains, ranging from information seeking and online disclosure to misinformation detection and content moderation. The presentation will discuss the pitfalls of heuristic-driven information processing and suggest solutions derived from the speaker's model of Human-AI-interaction based on his Theory of Interactive Media Effects (HAII-TIME) for promoting positive machine heuristic and undertrust caused by negative machine heuristic. Allowing media users to interact with AI-based media can not only result in better calibration of trust in AI media but also lead to more informed massessment of the credibility of AI-generated and AI-mediated information.

Generative AI: Biased but Trusted?

Markus Appel | University of Würzburg; Tanja Messingschlager | University of Würzburg

The proliferation of generative AI has come with new opportunities but also with substantive challenges to individuals and societies. The focus in this presentation is on algorithmic bias, the challenges that texts and pictures created with generative AI can be heavily biased against women and minorities, by neglecting them or by portraying them in stereotyped roles. At the same time, humans seem to trust AI more than other humans to portray the world in an unbiased way: People seem to apply a machine heuristic that leads to more trust than warranted. We examined algorithmic bias in a series of studies. We first showed (among other things) that text-to-image (T2I) generative AI platforms underrepresent women when asked for ‘a portrait of a competent person’ as compared to ‘a portrait of a person’ with one popular platform generating no images of women in the former condition (but generating a majority of images showing women in the latter condition). Following this content analysis, we examined individuals’ awareness of the algorithmic bias phenomenon, including its origins and consequences, in in-depth interviews and in a quantitative survey. Based on the insights gained in all three studies, potential countermeasures are dicussed.

Trusting the Machine? How AI Authorship Shapes News Consumption

Daria Kravets-Meinke | bidt; Hannah Schmid-Petri | University of Passau & bidt

News organizations worldwide are increasingly integrating artificial intelligence (AI) into multiple stages of the news production process. This development raises an important question: How do audiences respond to AI in journalism, particularly when AI is explicitly presented as the (co-)author of news content.
In a series of experimental studies, we investigate the effects of different authorship cues (human, AI, and AI-assisted) on users’ likelihood of selecting news items. We further analyze whether perceived credibility of the author mediates these effects, and how individual attitudes toward generative AI–conceptualized as a dispositional machine heuristic–influence these relationships. Results indicate that the machine heuristic plays a central role in shaping both credibility perceptions and selection behavior for AI-generated news. Building on these findings, we conduct a secondary analysis to examine the antecedents of the machine heuristic in greater detail. Specifically, we investigate its mediating role in the relationship between media trust, political orientation, populist attitudes, use of generative AI, AI literacy, fear of AI, age, education, and gender, and individuals’ credibility perceptions of a presumably AI- generated journalistic message. Finally, we address conceptual and methodological challenges in measuring the machine heuristic and present initial results from a pilot study employing an Implicit Association Test to capture underlying cognitive processes.

SUB-SESSION

A Digital Future Worth Living: From Ethical Principles to Practice in Software Engineering

19.11.2026, 15:30–17:00

Language: English

Chairs: Niina Zuber | bidt; Klaus Staudacher | bidt; Erich Prem | eutema, TU Vienna

Digital transformation research has generated a rich body of ethical principles, including fairness, autonomy, transparency, and accountability. Yet a persistent challenge remains: how can these principles be translated into everyday software engineering practice? At the same time, current debates on AI ethics are often dominated by risk mitigation and regulation, while offering only limited perspectives on what a desirable digital future should look like.

This roundtable brings these two questions together. It explores how ethical reflection can move beyond abstract principles and become embedded in software development processes, while also asking what kind of digital future is worth striving for. Drawing on the paradigm of Digital Humanism, the session examines how software systems can support human agency, responsibility, and flourishing rather than merely ensuring compliance.
The discussion combines philosophical perspectives with practical approaches from software engineering. It addresses questions such as: How can ethical values be operationalized in agile development? How can interdisciplinary collaboration between philosophy and engineering improve technology design? And how can software contribute to a digital society that promotes human dignity, meaningful work, trust, and social responsibility?

The roundtable brings together leading researchers from philosophy, AI ethics, and software engineering to discuss approaches such as Embedded Ethics, Value-Sensitive Design, and ethical deliberation in agile development. Contributions draw on virtue ethics, deontological ethics, and care ethics to develop richer normative perspectives for digital transformation.
Designed as an interactive exchange between speakers and participants, the session invites researchers and practitioners interested in responsible software engineering, AI ethics, Digital Humanism, and the future of digital society to discuss how ethical principles can become a lived practice in the design and development of digital technologies.

Impulse Statements & Roundtable Discussion:

Introduction

Klaus Staudacher | bidt & European Digital Humanism Initiative

Software is now embedded in almost all areas relevant to society, yet it is not value-neutral, and the guidelines and codes of ethics intended to address this issue remain often abstract and are limited in their effectiveness. Digital Humanism can contribute in several ways to overcoming this principle–practice gap.
Technology doesn't shape society on its own – people do. Digital Humanism is the idea that we, as individuals and societies, can and must actively shape the digital world from a humanistic perspective. For Digital Humanism, the question of the conditions for a good life in the digital age is fundamentally a question of the possibilities for the development of human authorship: How can we design digital technologies and their applications in ways that enable human authorship and flourishing for the benefit of all?

Digital Humanism is both a theoretical position and an ethics for the digital age.
From a theoretical point of view, it rejects positions such as the mechanistic paradigm (“humans are machines”) and the animistic paradigm (“machines are (like) humans”) that question or blur the line between humans and machines. It insists therefore that ethical and legal responsibility must remain with human agents rather than being delegated to AI systems and that ethically relevant decisions should and can not be made by AI. From a practical point of view it formulates a human-centred approach with principles that put people first by placing human dignity, democratic governance, social cohesion, and environmental sustainability at the centre of technological development.

Beyond AI Ethics: What Kind of Digital Future Do We Want? From Risk Mitigation to Human Flourishing

Erich Prem | eutema, TU Vienna & European Digital Humanism Initiative

Over the past decade, AI ethics has been remarkably successful in identifying ethical principles and reducing technological risks through concepts such as fairness, transparency, accountability, and trustworthiness. Yet these approaches primarily focus on preventing harm rather than articulating a positive vision of the digital future. This opening discussion asks whether Digital Humanism offers more than another ethical framework by providing a normative vision of technologies that actively promote human flourishing, democratic participation, and meaningful social practices.

From Principles to Practice: Can Digital Humanism Become a Design Paradigm? Embedding Human Values into Software Engineering

Alexander Pretschner | bidt & TU Munich

If Digital Humanism is to become more than a philosophical vision, its values must be translated into concrete design decisions, engineering methods, and organizational practices. This discussion explores how ethical reflection can be integrated into software engineering through approaches such as Value Sensitive Design, Embedded Ethics, and responsible AI development. Particular attention will be given to interdisciplinary collaboration between philosophers, software engineers, designers, and decision-makers, as well as to methods for translating abstract values into actionable requirements throughout the development lifecycle.

The Limits of AI Engineering: What Cannot Be Designed? Human Judgment, Responsibility, and Democratic Governance

Steven Umbrello | Institute for Ethics and Emerging Technologies, University of Turin & European Digital Humanism Initiative

While many ethical requirements can be operationalized in software, fundamental questions remain about the limits of technological design. Can responsibility, practical wisdom, democratic legitimacy, or human agency be embedded into AI systems, or do these ultimately depend on human institutions and social practices? The concluding discussion examines the relationship between engineering, governance, and human responsibility, asking where technical solutions end and where human judgment and democratic oversight must remain indispensable.

bidt Digital Transformation Research Conference

Sessions
Day 2

Friday, 20.11.2026

ECONOMY & WORK Session 3

Governing Digital Transformation Under Tensions: Between Regulation, Organizational Dynamics, and New Dependencies

20.11.2026, 09:00-10:30

Language: English

Programme Committee: Angela Graf | bidt; Sabine Pfeiffer | bidt & FAU Erlangen-Nuremberg; Thomas Hess | bidt & LMU Munich; Alexander Pretschner | bidt & TU Munich

Session Chair: Sabine Pfeiffer | bidt & FAU Erlangen-Nuremberg

Regulating BigTech: YouTube's Response to Australia's Social Media Ban

Ina Partisch | Fraunhofer FIT / Frankfurt University of Applied Sciences; Michael Rosemann | Queensland University of Technology, Australia; Nadine Ostern | Queensland University of Technology, Australia

Contextual Setting

On 10 December 2025, Australia became the first country to enforce a minimum age requirement for major social media platforms. The so-called Social Media Minimum Age (SMMA) law requires providers of ten selected platforms to take "reasonable steps" to ensure that users under the age of 16 cannot create or maintain accounts. The policy is widely regarded as a global test case and is being closely observed by regulators internationally. For example, European Commission President Ursula von der Leyen described Australia's approach as a "world-first" and "world-leading" model from which Europe could learn, while several European countries, including Germany, have intensified discussions about stricter age restrictions and age verification mechanisms for social media platforms. Underage users who continue to access age-restricted platforms are not penalised, nor are their parents or carers. Instead, the legislation is accompanied by substantial civil penalties of up to AUS$49.5 million for social media platforms that fail to comply. The SMMA law reflects a significant shift in the regulation of digital platforms. Rather than primarily attempting to regulate harmful content (Chatterjee et al., 2026; DeNardis & Hackl, 2015) or platform behaviour (Balkin, 2021; Zhang et al., 2026), the legislation focuses on regulating access to the platforms themselves. In doing so, regulation is shifted away from the product and toward the user. This is distinctively different to earlier regulatory approaches centred on content moderation through platform self-regulation (Balkin, 2021; Birkinshaw, 2024; Kokshagina et al., 2023) as it makes access to the platform environment itself the central governance issue.

Usually, Governments face several challenges when regulating Big Tech (Flew & Gillett, 2021). Social media platform companies are known as powerful forces that use their extraordinary economic, social, and political influence to shape regulation (Kane et al., 2025; Kokshagina et al., 2023). When confronted with regulatory threats, Big Tech companies often seek to influence regulatory outcomes by challenging the legitimacy of proposed laws, threatening, and mobilising public support. Through ongoing institutional work, platform companies actively participate in shaping regulatory requirements rather than merely complying with them (Kokshagina et al., 2023). Governments and firms continuously have to adapt to each other's actions, creating a dynamic regulatory environment characterized by strategic. But unlike many previous forms of digital regulation, the SMMA law constrained platforms' ability to rely on established response strategies (Birkinshaw, 2024). The combination of a clear policy objective combined with an accelerated implementation period while leaving specific compliance rules unspecified significantly challenged the social media platforms in their response strategies and therefore triggered a conflict situation between the parties, in which the platforms' responses escalated gradually over time (Baldwin et al., 2011; Pondy, 1967).

Research Problem and Research Question

Prior studies examine how platform firms interpret and comply with regulatory requirements or how digital platforms respond strategically to regulatory pressures (Birkinshaw, 2024; Chatterjee et al., 2026; Kokshagina et al., 2023). They also examine how regulatory frameworks themselves emerge when actors attempt to define and govern new technological phenomena or create incentives for platforms to act responsibly (Balkin, 2021; Seidel et al., 2025). While there is already an extensive body of literature regarding the regulation of Big Tech, less attention has been paid to how platform responses evolve throughout the implementation of a regulatory framework that combines an accelerated timeline and unspecified compliance requirements with the unique aim of user base exclusion. In particular, there is limited knowledge on how the inherent conflict between the two parties involved in the regulation is mitigated in light of the accelerated, unspecified conditions. Against this background, the study asks: *How do social media platform companies such as YouTube respond to Australia's Social Media Ban?*

Related Literature and Relevant Theoretical Lenses

To examine platform responses to the implementation of the SMMA law, the paper draws on literature concerning regulating Big Tech and platform responses to regulation. A comprehensive literature review identified conflict escalation theory as the most adequate theoretical lens for this research as the SMMA law represents a weakening of the platform's power position and thus lets them perceive themselves to be threatened by and in conflict with the regulators (Pondy, 1967). Rather than evaluating the merits of the legislation itself (Barnes et al., 2026; Finlay & Hooton, 2026), this study focuses on the implementation process, the mechanisms through which the law was operationalised, and how the strategic responses of a platform, can be interpreted from a theoretical lens allowing the study of escalation and conflict mitigation. By applying phenomenon-based theorizing to study the mechanisms behind the distinctive implementation period of the SMMA law, this paper provides a platform for bridging academic research and management practice. The paper aims to reconstruct the implementation of the SMMA law, analyse YouTube's public responses to it and examine with the lens of conflict escalation theory how those responses evolved in light of acceleration and underspecified compliance rules.

Research Design

This study adopts an exploratory qualitative single-case research design to examine how a major social media platform, YouTube, responds to accelerated and unspecified regulation. Despite several platforms affected by the SMMA law, this study focuses on Google's platform YouTube as the focal platform as it occupied a particularly contested position within the policy debate because of its controversially discussed affordances in learning and education (Balkin, 2021). Furthermore, YouTube engaged extensively in public communication and interacted intensively with the government during the implementation period, because it was initially exempted from the legislation before later being included, which provides a rich empirical record of its interactions with regulators. Methodologically, the study draws on qualitative document analysis while borrowing analytical strategies for developing theory from processual
data to reconstruct the sequence of regulatory events, activities, and implementation decisions that shaped the evolution of the SMMA law and to examine how YouTube strategically responded to these developments over time (Black, 2002).

Results To Date

The reconstruction of the implementation timeline of the SMMA law reveals the three major phases development of law, development of guidance, and commencement. Building on that, YouTube's responses to the implementation are analysed using a one-sided lens of conflict escalation theory and translated into a process framework, in which the key regulatory trigger events of the implementation process, their outcomes, and YouTube's responses are triangulated through cross-document analyses. This process framework explains

the response strategy of YouTube as an escalating conflict containing of prevention, followed by coercion, and ending with destruction.

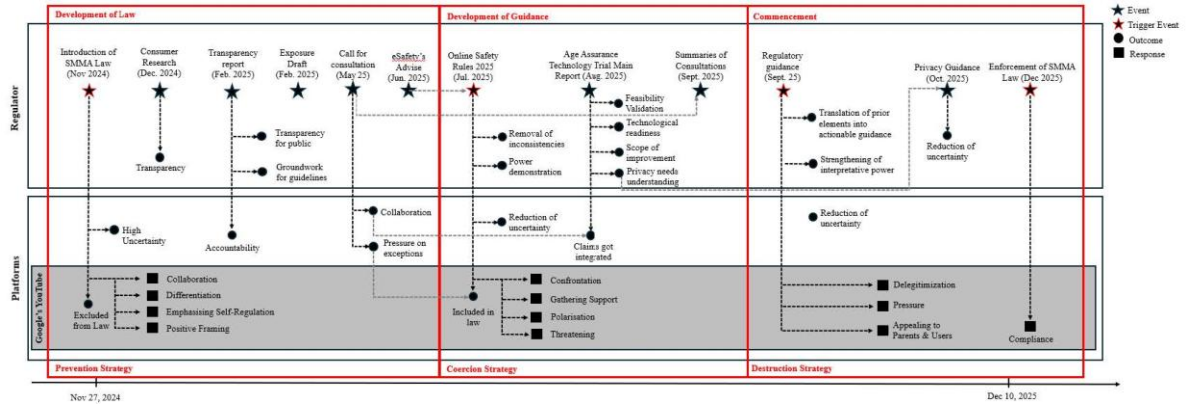


Figure 1: Process Model of YouTube's Responses to the Implementation of the SMMA Law

First, the platform adopted a prevention strategy by emphasising collaboration, differentiation, self-regulation, and positive framing in its public statements, with the aim of eliminating the perceived threat. During this phase, the government considered exempting YouTube from the law and preserving access to the platform. Following consultations with key stakeholders, the government published the updated Online Safety Rules in July 2025, which included YouTube in the restrictions. This not only shifted the broader implementation process from the initial development of the law toward regulatory specification through the development of guidance but also marked a clear change in YouTube's public response strategy. The second response strategy is characterised by coercive confrontation, gathering support, polarization, and threatening. During this phase, the platform also approached parents and users with their statements, aiming to gain sympathy for its position while simultaneously increasing pressure (Glasl, 1982). Following the consultation process and age assurance trial, the government published final regulatory guidance, marking the commencement of the law. In this phase, YouTube transitioned to a more destructive response strategy. By repeatedly labelling the law and its implementation as "rushed", "not fulfilling its promise", and "failed to allow for adequate consideration of the real complexities", YouTube increasingly sought to undermine the legitimacy and sanction potential of the regulatory approach (Glasl, 1982). This phase is characterized by delegitimization, pressure, and appealing to parents and users, further drawing them into the conflict. The process model suggests that platform responses evolve from cooperative and supportive positioning toward increasingly confrontational forms of institutional work, which aligns with previous studies on strategic platform responses to regulation (Kokshagina et al., 2023). Furthermore, based on the findings in the SMMA context, the paper proposes a generalized process model for one-sided conflict escalation under accelerated and unspecified regulation.

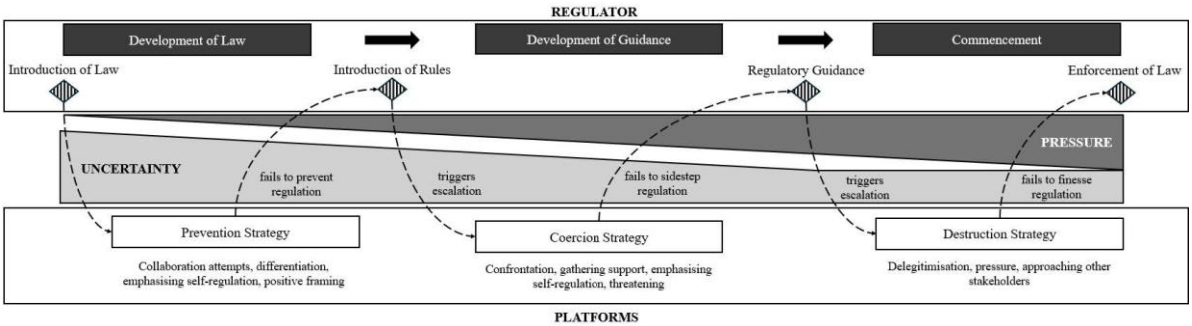


Figure 2: Process Model of One-Sided Conflict Escalation under Accelerated and Unspecified Regulation

Overall, the results suggest that accelerated implementation combined with residual regulatory uncertainty progressively reduces opportunities for strategic responses.

Expected Contributions

This study contributes to research on digital regulation in three ways. First, it examines how platform responses evolve over time under conditions of accelerated and unspecified regulation. Second, it introduces conflict escalation theory as a lens for understanding platform-regulator interactions during regulatory implementation. Third, it provides one of the first detailed analyses of the implementation of Australia's SMMA law. As an exploratory qualitative case study, the paper seeks to contribute to a deeper understanding of how a major platform company responds to accelerated and unspecified digital regulation. The resulting framework captures how the implementation of the SMMA law unfolds and how responses by platforms change over time in such an unprecedented digital regulation case. While this study focuses specifically on YouTube's strategic responses during the implementation phase of the SMMA law, the findings may also provide a foundation for broader future research on state-platform power relations and emerging forms of digital governance. Future research could build on this case to further examine how regulatory mechanisms shape negotiation dynamics between states and platform companies, and how power is exercised, challenged, and redistributed throughout regulatory processes.

The study contributes to the session "Economy and Work" by examining how a major platform company responds to the world's first comprehensive law placing the onus on digital exclusion for social good. Rather than unfolding linearly, the implementation process was characterised by continuous interaction and negotiation between governmental actors, platform companies, and non-governmental actors such as civil society. Governance therefore emerged as a dynamic process shaped by ongoing adaptation to regulatory trigger events and shifting institutional expectations. Similar to broader digital transformation processes, the governance of social media platforms requires constant organisational adjustment and strategic repositioning. While a comparatively stabilised state of compliance can be identified at the point of enforcement, the study seeks to reconstruct how this state emerged and which strategic responses were used to negotiate, delay, frame, or influence the implementation process over time. By combining perspectives from Information Systems research, digital governance, strategic communication, and organisational theory, the paper contributes to the session's interdisciplinary discussion of governance processes in the digital transformation of economy and work.

References

Balkin, J. M. (2021). How to Regulate (and Not Regulate) Social Media Symposium: Free Speech and Social Media Platform Regulation. Journal of Free Speech Law, 1(1), 71–96.

Barnes, C., Hall, A., Mantach, S., Oldmeadow, C., Attia, J., Backholer, K., Williams, C., Tefera, Y., Kay, F., & Wolfenden, L. (2026). Assessing early effects of Australia's Social Media Minimum Age Act on adolescents' social media use: Observational study. BMJ, 393, e363695. https://doi.org/10.1136/bmj-2026-363695

Birkinshaw, J. (2024). Too Little, Too Late? How Policymakers and Regulators Respond to the Business Model Innovations of Digital Firms. Academy of Management Perspectives.

Black, J. (2002). Regulatory Conversations. Journal of Law and Society, 29(1), 163–196. https://doi.org/10.1111/1467-6478.00215

Bodtker, A. M., & Katz Jameson, J. (2001Emotion in Conflict Formation and its Transformation: Application to Organizational Conflict Management. International Journal of Conflict Management, 12(3), 259–275. https://doi.org/10.1108/eb022858

Chatterjee, T., Gupta, A., & Thomas, P. (2026). One Size Doesn't Fit All: Understanding Social Media Platform Regulation in India. Communications of the Association for Information Systems, 58(1). https://aisel.aisnet.org/cais/vol58/iss1/71

DeNardis, L., & Hackl, A. M. (2015). Internet governance by social media platforms. Telecommunications Policy, SPECIAL ISSUE ON THE GOVERNANCE OF SOCIAL MEDIA, 39(9), 761–770. https://doi.org/10.1016/j.telpol.2015.04.003

Finlay, L., & Hooton, P. (2026). Fast law, fragile rights: The human rights consequences of Australia's rapid social media reforms. Griffith Journal of Law & Human Dignity, 13(1). https://doi.org/10.69970/gjlhd.v13i1.1288
Flew, T., & Gillett, R. (2021). Platform policy: Evaluating different responses to the challenges of platform power. Journal of Digital Media & Policy, 12(Platform Governance: Power, Diversity and Accountability), 231–246. https://doi.org/10.1386/jdmp_00061_1
Glasl, F. (1982). The Process of Conflict Escalation and Roles of Third Parties. In: Bomers, G.B.J., Peterson, R.B. (eds) Conflict Management and Industrial Relations. Springer, Dordrecht. 119–140.
Kane, G.C. et al. (2025). The Evolution of Social Media Networks : Revisiting and Extending Kane et al. (2014): 10 Years Later, MIS Quarterly, 49(4), ix-xvi.
Kokshagina, O., Reinecke, P. C. & Karansios, S. (2023). To regulate or not to regulate: unravelling institutional tussles around the regulation of algorithmic control of digital platforms. Journal of Information Technology, 38(2), 160-179.
Pondy, L. R. (1967). Organizational Conflict: Concepts and Models. Administrative Science Quarterly, 12(2), 296–320. https://doi.org/10.2307/2391553
Seidel, S., Frick, C. J., vom Brocke, J. (2025). Regulating Emerging Technologies: Prospective Sensemaking Through Abstraction and Elaboration. MIS Quarterly, 49(1), 179-204.
Zhang, X., Wei, Z., Du, Q., & Zhang, Z. (2026).Social Media Moderation and Content Generation: Evidence From User Bans1. MIS Quarterly 1 March 2026; 50 (1): 211–242.

Are We Now Being Surveilled? Organizational Ignorance in Digitalized Waste Management

Alena Bleicher | Harz University of Applied Sciences; Sabine Biedermann Camposano | Harz University of Applied Sciences & TU Munich; Diana Ayeh | Helmholtz Centre for Environmental Research

Uncertainties in production and administration caused by knowledge gaps and a lack of information is a basic problem of organizations. Therefore, reducing uncertainties by eliminating knowledge gaps is a central focus of organizations and their management (Dewett and Jones 2001; Schaller 2016). In this regard, digital innovations promise to generate better insights into the organization and its markets, and make external knowledge available, facilitate organizational memory, prevent human error and uncover knowledge gaps through the integration of data (Roberts 2023).

Despite the positive effects attributed to information and knowledge sharing by the help of digital technologies, references can be found in the literature that question both the potential of technologies in terms of these expectations and the effects of transparency. They even suggest that digital technologies may have the opposite effect and lead to less knowledge rather than more (Büchner 2018a). Moreover, new possibilities of knowledge provided by digital technologies are not always desirable. Rather, technology users in organizations may intentionally create blind spots e.g. in order to strengthen (power) positions (Mormann 2016; Büchner 2018b; Heinlein and Huchler 2023).

While research on the aspects and effects of digitalization in organizations focused on increasing transparency, information flows, and knowledge gains their flipside – ignorance in various forms – is disregarded so far. This is where the contribution of this paper comes in. Based on an empirical study in the German waste management sector we investigated the question: To what extent digitalization in organizations and ignorance mutually shape?

In order to study this question, the contribution at hand relies on conceptual ideas from the field of organizational ignorance. Research in this field has revealed the relevance of ignorance – in forms such as knowledge gaps, errors, secrets, strategic ignorance, or categorial blindness – for shaping organizational

processes and decision-making (Roberts 2023; Jalonen 2024). As such perspective emphasizes, ignorance cannot always be easily transformed into knowledge by acquiring more knowledge. Rather, ignorance is the result of social construction in which individual attitudes are as important as interpersonal relationships and organizational structures and cultures (Jalonen 2024). Ignorance may be an asset , e.g. with regard to creativity (Ferraro et al. 2015; Roberts 2023). It can be strategically used or willfully sustained in order to ensure stability of and power positions within an organization (McGoey 2012; Bakken and Wiik 2018; Alvesson et al. 2022). We studied the mutual shape of digitalization and ignorance based on an ethnographic study in two waste management organizations in Germany. Digitalization in waste management organizations comes with the promise of creating more and better knowledge on waste (Strzelecka 2025) and on organizational processes related to waste collection and treatment (i.e. logistics). Both are considered a precondition for more efficient organizational processes and more informed waste management decisions. In the moment of our research in both case studies digitalization has been an ongoing and incomplete process, marked by multiple uncertainties for those involved in its planning and implementation.

Our results show how organizational ignorance and digitalization processes mutually shape. The promise of transparency offered by digital technologies necessitates the continuous renegotiation of secrets, i.e., who is allowed to know what within the organizations, in daily work across hierarchies. We identified administrative measures, technological solutions, and individual strategies that are arranged to govern the transparency-secrecy tension. We further revealed how individuals strategically resort to ignorance in order to mitigate the negative effects the introduction of digital technologies in the organization might have on them and their work. Understanding that ignorance is socially constructed and may have manifold effects in digitalization in organizations, allows for designing appropriate structures to govern digitalization.

Publication bibliography

Alvesson, Mats; Einola, Katja; Schaefer, Stephan M. (2022): Dynamics of wilful ignorance in organizations. In *The British journal of sociology* 73 (4), pp. 839–858. DOI: 10.1111/1468-4446.12963.

Bakken, Tore; Wiik, Eric Lawrence (2018): Ignorance and Organization Studies. In *Organization Studies* 39 (8), pp. 1109–1120. DOI: 10.1177/0170840617709312.

Büchner, Stefanie (2018a): Digitale Infrastrukturen - Spezifik, Relationalität und die Paradoxien von Wandel und Kontrolle.

Büchner, Stefanie (2018b): Zum Verhältnis von Digitalisierung und Organisation. In *Zeitschrift für Soziologie* 47 (5), pp. 332–348. DOI: 10.1515/zfsoz-2018-0121.

Dewett, Todd; Jones, Gareth R. (2001): The role of information technology in the organization: a review, model, and assessment. In *Journal of Management* 27 (3), pp. 313–346. DOI: 10.1016/S0149-2063(01)00094-0.

Ferraro, Fabrizio; Etzion, Dror; Gehman, Joel (2015): Tackling Grand Challenges Pragmatically: Robust Action Revisited. In *Organization Studies* 36 (3), pp. 363–390. DOI: 10.1177/0170840614563742.

Heinlein, Michael; Huchler, Norbert (2023): Artificial intelligence in the practice of work. In *WOLG* 17(1). DOI: 10.13169/workorgalaboglob.17.1.0034.

Jalonen, Harri (2024): Ignorance in organisations – a systematic literature review. In *Manag Rev Q* 74, pp. 909–950. DOI: 10.1007/s11301-023-00321-z.

McGoey, Linsey (2012): Strategic unknowns: towards a sociology of ignorance. In *Economy and Society* 41 (1), pp. 1–16. DOI: 10.1080/03085147.2011.637330.

Mormann, Hannah (2016): Das Projekt SAP. Zur Organisationssoziologie betriebswirtschaftlicher Standardsoftware. 1st ed. Bielefeld: transcript (Sozialtheorie). Available online at https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=4974764.

Roberts, Joanne (2023): Organizational Ignorance. In Matthias Groß, Linsey McGoey (Eds.): Routledge International Handbook of Ignorance Studies. second edition. London, New York: Routledge Taylor & Francis, pp. 367–376.

Schaller, Philipp D. (2016): Experimentelle Organisationsentwicklung: Nomos Verlagsgesellschaft GmbH & Co. KG.
Strzelecka, Celina (2025): Green-digital transition in municipal waste management: Ethnographic perspective on 'smart' waste monitoring and management system. In *Environmental Science & Policy* 167, p. 104037. DOI: 10.1016/j.envsci.2025.104037.

Beyond Individual Overreliance: Generative AI Dependence, Workplace Norms and the Governance of Everyday Automation

Christoph Lutz | BI Norwegian Business School, Norway; Gemma Newlands | University of Zurich, Switzerland; Shubin Yu | HEC Paris, France

Artificial intelligence (AI) is increasingly embedded in routine work and everyday life. Public and policy debates often approach this development through the language of trustworthy AI, emphasizing fairness, transparency, accuracy, privacy and reliability as preconditions for legitimate adoption (HLEG, 2019; Kaur et al., 2022). This perspective is important but can obscure a pressing issue: AI systems may be used too readily, too habitually, or in contexts where their outputs should remain provisional. Appropriate reliance requires calibration between system capability, task context and human judgment (Lee & See, 2004).
Research on AI overreliance has expanded rapidly in HCI and related fields, often through experiments or design interventions that seek to reduce overtrust in decision-support systems (Bo et al., 2025; Buçinca et al., 2021; Klingbeil et al., 2024). Much of this work focuses on bounded tasks or high-risk domains. Less is known about how people understand overreliance in mundane settings: navigation, customer service, social media feeds, workplace writing, teaching, creative work and generative AI use. In such settings, AI adoption is often pragmatic and driven by external pressures rather than in completely thoughtful and rational way. People use AI because it is convenient, fast, easy, normalized or difficult to avoid, not necessarily because they have evaluated whether a system is trustworthy (Heuser & Vulpius, 2024; Wu et al., 2025).
Our research examines perceived overreliance on AI in everyday and work-related contexts. It treats overreliance as a socially embedded pattern shaped by convenience, habit, professional routines, institutional substitution and changing norms around competence. In doing so, we speak directly to questions of digital transformation in economy and work: how AI reorganizes practices, shifts responsibility, creates new dependencies and raises questions about skills, autonomy and institutional governance.

Research question and conceptual orientation
The central research question is: How do people perceive the forms, drivers and consequences of AI overreliance in daily life and work? We focus on overreliance rather than overtrust because reliance is closer to observable practice and avoids some of the anthropocentric assumptions built into the concept of trust (Freiman, 2023; Klingbeil et al., 2024).
Conceptually, we combine three perspectives. First, trust calibration research highlights that both underuse and overuse of AI can be problematic when reliance is poorly matched to system performance (Lee & See, 2004). Second, domestication theory helps explain how technologies become normalized and move into the background of everyday life, where they are used with little conscious scrutiny (Silverstone & Haddon, 1996; Brause & Blank, 2020). Third, research on cognitive frugality suggests
that users often conserve effort by relying on heuristics and default options, particularly when tools make a task easier or less cognitively demanding (Fiske & Taylor, 2020; Gerlich, 2025). The three perspectives shift the focus from isolated user errors to the routines and contexts in which reliance becomes convenient, habitual and difficult to question.

Method

We draw on an online survey conducted in the United Kingdom in August 2024. Participants were recruited through Prolific using its representative sample option for age, gender and ethnicity. The final sample consisted of 292 respondents. Gender was balanced (50.9% women, 48.8% men, 0.3% neither male nor female), age ranged from 18 to 77 years (M = 46.3, SD = 15.5), and respondents had a broad range of educational backgrounds.

The survey combined closed and open-ended questions. Closed-ended measures covered general opinion toward AI, self-reported AI literacy (Wang et al., 2023), attitudes toward AI (Sindermann et al., 2021), frequency of use of eight AI technologies, and perceived societal overreliance on those same technologies. The eight technologies were text-generation tools, image-generation tools, chatbots, social media algorithms, smart speakers, mapping applications, video-game AI and social robots. Open-ended questions asked respondents to describe instances of overreliance, reasons for overreliance, consequences of overreliance, and why people in the UK rely too little or too much on specific AI technologies.

The quantitative data were analyzed descriptively and through correlations. The qualitative responses were analyzed with thematic coding (Braun & Clarke, 2006). Codes were grouped around the survey structure: instances and forms of overreliance; antecedents and reasons; and perceived outcomes and consequences. The aim was an empirically grounded account of how users describe AI overreliance across everyday and occupational contexts.

Findings

Table 1. Perceived societal overreliance on AI technologies in the UK

AI technology	Mean	Interpretation
Social media algorithms	59.42	Highest perceived overreliance
Mapping applications	47.49	High routine dependence
Smart speakers	37.73	Moderate overreliance
Chatbots	37.25	Moderate overreliance
Text-generation tools	33.09	Moderate overreliance
Video-game AI	23.25	Lower concern
Image-generation tools	11.84	Limited perceived overreliance
Social robots	-6.44	Perceived underuse rather than overuse

Respondents perceived the strongest societal overreliance on AI in social media, followed by mapping applications, smart speakers, chatbots and text-generation tools. Social robots were the only technology for which the mean was negative, suggesting perceived underuse rather than overuse. The pattern is notable because the technologies associated with overreliance are not necessarily the newest or most spectacular systems but represent tools deeply woven into daily routines: recommendation feeds, navigation systems, customer-service interfaces and writing assistants.

Qualitative responses showed that overreliance was frequently narrated through specific work and everyday practices. Mapping applications were often described as making people less attentive to their surroundings and less willing to plan routes independently. Generative AI tools were linked to teaching, writing, coding, design and job applications. One university tutor described using ChatGPT to generate multiple-choice questions and later discovering that some answers were incorrect or ambiguous. A design-oriented respondent described using AI-generated architectural sketches for early concept development, while also worrying that the outputs subtly influenced their thinking despite being technically weak. These accounts suggest that overreliance often emerges where AI saves time but also shapes judgment, standards and creative direction.

Several respondents located overreliance in organizations. Customer-service chatbots were a recurring example. Respondents described companies as replacing accessible human support with automated systems

that could not handle complex queries. This finding broadens the object of analysis. Overreliance is not only an individual tendency to believe an AI output but can also be an organizational practice in which institutions shift work, responsibility and friction onto automated systems, with consequences for service quality and user trust.

The most commonly mentioned reasons for overreliance were laziness or complacency, time saving, ease of use, convenience and perceived performance. Other reasons included habit, lack of confidence, lack of skills, lack of alternatives, curiosity, novelty and media hype. These accounts were often morally charged, but they also revealed practical pressures. Respondents described AI as attractive because it made unpleasant or demanding tasks easier, reduced uncertainty, and offered a ready-made path through cognitive work. In workplace contexts, such pressures may be intensified by productivity expectations and the growing sense that competent professionals should know how to use AI.
The main perceived consequences were deskilling, misinformation and diminished critical thinking. Respondents also mentioned time loss, job displacement, frustration, loss of creativity, reduced originality and weakening of professional identity. Some accounts were ambivalent. AI was seen as useful, productive and sometimes necessary, while carrying risks of cumulative dependence. One respondent described relying on text-generation tools for job application letters, only to find that the outputs were too generic and less successful than self-written applications. Another respondent reflected that AI-generated work no longer matched their usual style and might degrade artistic skills through lack of practice. These examples point to subtle harms that may unfold over time rather than dramatic failures.

Contribution to economy and work research
The research contributes to research on digital transformation in economy and work in several ways. First, it extends research on AI overreliance beyond high-risk and laboratory settings by showing how dependence is perceived in everyday and occupational practices. Second, it highlights the multi-level character of overreliance. Individuals may rely on AI for navigation, writing, teaching or creative work, while organizations may rely on AI to restructure customer service, reduce labour costs or standardize interaction. Third, it connects overreliance to questions of competence, professional identity and work quality. Respondents consistently worried that people might lose skills, confidence, originality and the ability to judge when automation is appropriate.

The findings also complicate simple design remedies. Friction, explanations and warnings may reduce inappropriate reliance in some settings, but they can also create dissatisfaction or even increase trust in incorrect recommendations (Bo et al., 2025; Passi & Vorvoreanu, 2022). A governance-oriented approach therefore needs to consider work norms, institutional incentives and domain-specific expectations. The relevant question is under which social, professional and organizational conditions reliance remains proportionate, reviewable and accountable.

Implications and limitations
For policy and practice, our study points to the need for AI literacy that goes beyond functional competence. Users need resources for assessing when AI is useful and when human judgment is indispensable. Workplaces need clearer norms about acceptable AI reliance across tasks. Such norms should address not only accuracy, but also responsibility, disclosure, skill development and the preservation of meaningful human expertise.
Our study has certain limitations. The data are self-reported and capture perceptions rather than observed behaviour. The UK sample allows a broad view of everyday interpretations, but cross-national work would allow to examine how overreliance varies across cultures and regulatory environments. The survey format also limits the depth of individual accounts. Future research should combine surveys with interviews, observational studies and behavioural data to examine how AI reliance develops over time, especially as generative AI becomes further integrated into workplace tools.

References

Bo, J. Y., Wan, S., & Anderson, A. (2025). To rely or not to rely? Evaluating interventions for appropriate reliance on large language models. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (pp. 1-23).

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101.

Brause, S. R., & Blank, G. (2020). Externalized domestication: Smart speaker assistants, networks and domestication theory. Information, Communication & Society, 23(5), 751-763.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1-21.

Fiske, S. T., & Taylor, S. E. (2020). Social cognition: From brains to culture (4th ed.). SAGE.

Freiman, O. (2023). Making sense of the conceptual nonsense 'trustworthy AI'. AI and Ethics, 3(4), 1351-1360.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies, 15(1), 6.

Heuser, M., & Vulpius, J. (2024). 'Grandma, tell that story about how to make napalm again': Exploring early adopters' collaborative domestication of generative AI. Convergence. https://doi.org/10.1177/13548565241285742

HLEG (EU High-Level Expert Group on AI). (2019). Ethics guidelines for trustworthy AI. European Commission.

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. ACM Computing Surveys, 55(2), 1-38.

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI: An experimental study on the extent and costs of overreliance on AI. Computers in Human Behavior, 160, 108352.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80.

Passi, S., & Vorvoreanu, M. (2022). Overreliance on AI literature review. Microsoft Research.

Palan, S., & Schitter, C. (2018). Prolific.ac: A subject pool for online experiments. Journal of Behavioral and Experimental Finance, 17, 22-27.

Silverstone, R., & Haddon, L. (1996). Design and the domestication of information and communication technologies. In R. Mansell & R. Silverstone (Eds.), Communication by design: The politics of information and communication technologies (pp. 44-74). Oxford University Press.

Sindermann, C., Sha, P., Zhou, M., Wernicke, J., Schmitt, H. S., Li, M., Sariyska, R., Stavrou, M., Becker, B., & Montag, C. (2021). Assessing the attitude towards artificial intelligence. KI - Künstliche Intelligenz, 35(1), 109-118.

Wang, B., Rau, P. L. P., & Yuan, T. (2023). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. Behaviour & Information Technology, 42(9), 1324-1337.

Wu, Q., Tian, J., & Liu, Z. (2025). Exploring the usage behavior of generative artificial intelligence: A case study of ChatGPT with insights into habit and personal innovativeness. Current Psychology. https://doi.org/10.1007/s12144-024-07193-w

LAW & ETHICS Session 3
Responsibility and Control in AI-Mediated Decision-Making

20.11.2026, 09:00-10:30

Language: English

Programme Committee: Niina Zuber | bidt; Julian Nida-Rümelin | bidt & LMU Munich; Dirk Heckmann | bidt & TU Munich; Felix Freiling | bidt & FAU Erlangen-Nuremberg

Session Chair: Johannes Reisinger | bidt & LMU Munich

Access to Digital Information as Infrastructure Relevant to Fundamental Rights – How Far Does State Responsibility Extend?

Charlotte Werther | Open Search Foundation e.V.

1. Einleitung

In der modernen Wissens- und Informationsgesellschaft ist Information ein Wirtschaftsgut, ein Machtfaktor und Kulturgut.

Der Beitrag geht der These nach, dass sich aus den Grundrechten eine staatliche Verantwortung für den unabhängigen, transparent strukturierten und funktionsfähigen Zugang zu digitalen Informationen ergibt, die über eine bloße Regulierung privater Anbieter hinausgeht. Muss der Staat gegebenenfalls selbst die notwendigen Infrastrukturen für den Informationszugang bereitstellen, um die grundrechtlichen Interessen seiner Bürger und Bürgerinnen zu schützen und Teilhabe und Wertschöpfung zu ermöglichen?

Die Organisation des Zugangs zu Informationen hat sich im Zuge der Digitalisierung grundlegend gewandelt. Eine der wesentlichen Informationsquellen heutzutage ist das Internet, das mehr und mehr andere Informationskanäle ersetzt. Dabei kommt Suchanwendungen, seien es die klassische Suchmaschine oder eine KI-gestützte Suche, eine Schlüsselrolle zu. Sie bestimmen durch Auswahl, Gewichtung und Präsentation von Informationen, welche Inhalte auffindbar sind und in welcher Form sie präsentiert werden. Suchanwendungen ermöglichen nicht nur den Zugang zu Informationen, sondern sie gestalten diesen Zugang aktiv.

Der Zugang zu Informationen im Internet ist aktuell geprägt durch datengetriebene Geschäftsmodelle und Quasi-Monopole außereuropäischer Technologiekonzerne. Die Technologiekonzerne sind in der Position, die Bezugs- und Nutzungsbedingungen zu diktieren. Dabei werden sie geleitet von außereuropäischen und kommerziellen Interessen. Daten, Know-How und Lizenzgebühren fließen ab, Wertschöpfung geht verloren. Und nicht nur das: Digitale Dienste und Infrastrukturen sind in den Mittelpunkt der Geopolitik gerückt. Deutschland und Europa sind erpressbar. Technologie wird als Druckmittel benutzt.

Während das Grundgesetz keinen Anspruch auf spezifische Inhalte garantiert, könnte es aber eine staatliche Pflicht zur Gewährleistung der strukturellen Voraussetzungen für einen unabhängigen, unvoreingenommenen Zugang zu transparent strukturierten Informationen begründen. Unabhängiger Zugang zu Informationen ist heute mehr denn je wichtig für Teilhabe, die Demokratie und Wertschöpfung.

2. Grundrechtliche Anforderungen und strukturelle Voraussetzungen

Die Bedeutung digitaler Informationen für Teilhabe, Wertschöpfung und die Demokratie wirft die Frage auf, wie der Zugang dauerhaft grundrechtskonform gestaltet werden kann und wer dafür verantwortlich ist.

Der Zugang zu Informationen ist die Grundvoraussetzung für die eigene Meinungsbildung und damit für eine freie und möglichst gut informierte demokratische Öffentlichkeit. Gemäß Art. 5 Abs. 1 Satz 1, 2. Alt. GG hat jeder das Recht, sich aus allgemein zugänglichen Quellen ungehindert zu unterrichten. Der Schutzbereich des

Grundrechts umfasst sowohl die passive Entgegennahme von Informationen als auch das Recht, sich Informationen selbst aktiv zu beschaffen. Dafür ist das Internet eine der wesentlichen Informationsquellen.
Um das Internet als Informationsquelle nutzen zu können und aus der Masse an Informationen relevantes herauszufiltern, braucht es Suchanwendungen. Ohne Suchanwendungen ist es nahezu unmöglich, Informationen im Internet zu finden. Die Grundlage für funktionsfähige Suchanwendungen sind strukturierte und ständig verfügbare Sammlungen qualitativ hochwertiger Webdaten, deren Kernelement ein sogenannter Webindex ist. Diese Webdateninfrastrukturen katalogisieren, gestalten und pflegen Informationen und Dokumente aus dem Internet thematisch, sprachlich, kulturell und inhaltlich. Dazu eingesetzte Auswahl- und Ranking-Algorithmen sind notwendigerweise selektiv, doch ihre Funktionsweise ist häufig intransparent. Wie die Webdateninfrastrukturen aufgebaut sind, welche Inhalte sie enthalten und wie diese Inhalte ausgespielt werden, ist für Dritte nicht nachvollziehbar. Unternehmen, die die Webdateninfrastrukturen gestalten, können somit beeinflussen, welche Informationen den Menschen zugänglich sind. Webdateninfrastrukturen werden im Wesentlichen von privatwirtschaftlich organisierten Unternehmen bereitgestellt, die außerhalb Europas operieren. Die bestehenden Webdateninfrastrukturen sind zudem beeinflusst von kommerziellen Interessen. Das führt zu Verzerrungen in der Informationsvermittlung und beeinträchtigt die Vielfalt der zugänglichen Informationen. Desinformation und Manipulation können die Folgen sein.
Diese strukturellen Probleme betreffen die Architektur des Informationszugangs und erfordern daher eine infrastrukturelle Perspektive.

3. Der offene Webindex als Modell einer grundrechtskonformen Infrastruktur

Um die Funktionsfähigkeit des Internets als unvoreingenommene Quelle digitaler Informationen zu erhalten, bedarf es alternativer, an den Grundrechten und –werten ausgerichteter offener Infrastrukturen, die alternative Suchanwendungen ermöglichen.
Den Prototypen einer offenen Webdateninfrastruktur hat ein Konsortium aus 14 europäischen Partnern entwickelt. Für diese Infrastruktur werden Webinhalte dezentral auf verschiedenen über Europa verteilten Rechenzentren gesammelt (gecrawlt), indexiert und gehostet. Ein offener Quellcode und offene Algorithmen machen die Funktionsweise des Systems transparent und nachvollziehbar. Durch öffentliche Moderation sollen die Auswahl und Struktur der indexierten Inhalte überwacht und so Vielfalt und Diskriminierungsfreiheit sichergestellt werden. Die gecrawlten und indexierten Datensätze sind anders als die Datensätze der großen Technologiekonzerne frei zugänglich und im Internet herunterladbar.
Auf der Ebene der Infrastruktur kann der offene Webindex im Wege des „values by design“ normative Anforderungen und Grundwerte verankern, z.B. im Hinblick auf Datenschutz und Fairness. Dazu gehören Datensparsamkeit bei der Zusammenstellung seiner Inhalte genauso wie transparente und diskriminierungsfreie Indexierungsverfahren und Rankingmechanismen, die nicht durch kommerzielle Interessen geleitet werden. Im Rahmen der Aufbereitung der Datensätze werden beispielsweise nicht erforderliche personenbezogene Daten herausgefiltert und nicht indexiert.
Der offene Webindex schafft so die Voraussetzung für einen pluralistischen und datenschutzfreundlichen Zugang zu Informationen. Die Trennung von Infrastruktur und Suchanwendung trägt dazu bei, dass unabhängige Anbieter auf der Grundlage des offenen Webindizes verschiedene innovative Suchdienste entwickeln können, ohne selbst einen vollständigen Index aufbauen zu müssen. Das reduziert Markteintrittsbarrieren und fördert den Wettbewerb.

4. Welche Verantwortung trägt der Staat?

Im Zusammenhang mit dem Aufbau und dem Betrieb einer Webdateninfrastruktur muss die Rolle des Staates betrachtet werden.
Für den Aufbau einer Webdateninfrastruktur werden erhebliche Ressourcen benötigt. Das Internet muss dauerhaft mit Hilfe von Bots durchsucht und eingesammelt (gecrawlt) werden.
Die gecrawlten Inhalte müssen aufbereitet und katalogisiert (indexiert) werden. Dafür werden große Rechenkapazitäten benötigt, die entsprechend teuer sind. Unabhängig davon, dass es vielen Unternehmen

schon an den erforderlichen Ressourcen fehlt, würde die Funktionsfähigkeit des Internets stark eingeschränkt werden, wenn viele Unternehmen eigenständig und parallel Webinhalte crawlen und aufbereiten. Intensives Content Scraping er Crawler kann zur Überlastung von Servern führen. Antwortzeiten beim Aufrufen von Webseiten verlängern sich.

Es stellt sich daher die Frage nach der staatlichen Verantwortung für den Aufbau und die Bereitstellung einer solchen Infrastruktur. Eine staatliche Verantwortung ist für verschiedene Bereiche anerkannt. Im Rahmen der Daseinsvorsorge hat der Staat die Aufgabe, die Bevölkerung und die Wirtschaft mit den als lebensnotwendig erachteten Gütern zu versorgen. Dazu gehören u.a. die Grundversorgung mit Energie und Wasser, der öffentliche Verkehr sowie Post- und Telekommunikationsdienste. Auch die Grundrechte können Schutz- und Gewährleistungspflichten begründen. In Bereichen wie dem Rundfunk- oder Telekommunikationsrecht hat sich daraus die Verpflichtung entwickelt, Rahmenbedingungen zu schaffen, die Vielfalt und Zugänglichkeit sichern. Das wirft die Frage auf, ob und inwieweit das Prinzip der Daseinsvorsorge und/oder die Grundrechte eine staatliche Verantwortung für den digitalen Informationszugang begründen. Eine Infrastrukturverantwortung kann sich aus der zentralen Bedeutung digitaler Informationen für Teilhabe, Wertschöpfung und die Grundrechte ergeben. Wenn die bestehenden Marktstrukturen diese nicht hinreichend gewährleisten, könnte der Staat verpflichtet sein, korrigierend einzugreifen.

5. Governance-Optionen

Der Status quo legt nahe, dass Regulierung allein nicht ausreicht, um den unabhängigen und vollumfänglichen Zugang zu Informationen hinreichend zu gewährleisten.21 Insbesondere die Fokussierung auf einzelne Suchdienste oder Akteure greift zu kurz, wenn zentrale Probleme schon auf der Ebene der Infrastruktur liegen. Vor diesem Hintergrund plädiert der Beitrag für eine institutionelle Absicherung durch mehr staatliche Verantwortung und aktive Gestaltung des Zugangs zu Informationen. Statt nur auf die Kontrolle einzelner Anbieter abzustellen, sollte der Staat aktiv Infrastrukturen schaffen, um so die Markteintrittsbarrieren zu reduzieren und den grundrechtskonformen Zugang zu digitalen Informationen zu ermöglichen. Verschiedene Governance-Strukturen sind dabei denkbar:

Ein staatliches Modell würde die direkte Verantwortung des Staates betonen, birgt jedoch Risiken im Hinblick auf politische Einflussnahme und Effizienz. Ein öffentlich-rechtliches Modell könnte eine größere Unabhängigkeit gewährleisten, ähnlich wie im öffentlich-rechtlichen Rundfunk. Hybride Public-Private-Modelle bieten die Möglichkeit, staatliche Zielsetzungen mit privatwirtschaftlicher Innovationskraft zu verbinden, werfen jedoch Fragen der Steuerung und Kontrolle auf.
Unabhängig vom konkreten Modell ist entscheidend, dass die Governance-Struktur geeignet ist, Souveränität, Vielfalt und Wettbewerb zu gewährleisten und so grundrechtliche Anforderungen dauerhaft zu sichern. Dazu gehören insbesondere Offenheit, Transparenz, Rechenschaftspflicht und die Möglichkeit externer Kontrolle. Gleichzeitig müssen Anreize für Innovation und Weiterentwicklung erhalten bleiben.

6. Fazit

Der Informationszugang ist zentrale Voraussetzung für eine funktionierende Informationsgesellschaft. Der derzeit ganz überwiegend privatwirtschaftlich organisierte und durch außereuropäische Monopole geprägte Zugang zu digitalen Informationen gefährdet Teilhabe, Wertschöpfung, Demokratie und die Grundrechte.
Der Beitrag entwickelt die These, dass der Zugang zu digitalen Informationen unter Bedingungen infrastruktureller Konzentration auch als staatliche Aufgabe verstanden werden muss. Dabei überträgt der Ansatz die Prinzipien der Daseinsvorsorge in den digitalen Bereich und versteht die Grundrechte nicht nur als individuelle Abwehrrechte, sondern als strukturierende Prinzipien für die Gestaltung zentraler technischer Systeme.

Der Beitrag leistet einen interdisziplinären Beitrag zur Debatte über Verantwortung, Vertrauen und Integrität, indem er die strukturellen Eigenschaften digitaler Infrastrukturen an der Schnittstelle Recht, Gesellschaft und

Technik betrachtet. Er richtet sich an Rechtswissenschaftlerinnen und Rechtswissenschaftler ebenso wie an Forschende aus Informatik, Politikwissenschaft und Science and Technology Studies, die an der Frage arbeiten, wie demokratische Gesellschaften die Kontrolle über ihre digitale Infrastruktur zurückgewinnen können.

Democratic Legitimacy under the Inferential Turn: Epistemic Power, Traceability Gaps, and Institutional Reconstructability in AI-mediated Public Administration

Thomas Absmayr | FHV - Vorarlberg University of Applied Sciences, Austria

I) The inferential turn and the transformation of justificatory authority

As artificial intelligence (AI) becomes increasingly embedded in public administration (PA), it is reshaping the epistemic conditions under which administrative authority is exercised (Covilla, 2025; Luise & Cocozza, 2025). AI systems affect not only administrative outputs, but also the evidentiary, classificatory, and justificatory architectures through which PA becomes institutionally intelligible (Wihbey, 2025; Zhu, 2025). Drawing on epistemic power theory (Archer et al., 2020), this contribution argues that AI-mediated public administration raises a significant yet underexplored challenge for democratic legitimacy (Binns, 2018; Díaz de la Cruz et al., 2025) – referring here not merely to procedural legality or administrative efficiency, but to the institutional capacity to exercise public authority through reconstructable, contestable, and normatively intelligible reasons (Beckman et al., 2024).

This challenge is referred to here as the inferential turn – a structural reconfiguration within public administration in which and through which probabilistic and classificatory infrastructures increasingly mediate administrative justification (Cohen, 2010). While PA has always relied on inference (Simon, 1997), this inferential turn is best understood not as a sudden rupture, but as a cumulative intensification crossing qualitative thresholds in scale, speed, and institutional distance from legally codified justificatory categories (Burrell, 2016; Criado et al., 2020; Oswald, 2018). This shift tends to trigger erosion of justificatory authority, understood here as the institutionally recognized and normatively defensible validity of reasons for administrative action (Boughey, 2021; Forst, 2012).

Against this background, this contribution introduces institutional epistemic reconstructability (IER) as the capacity to reconstitute, contest, and normatively anchor justificatory chains through which inferential outputs become authoritative administrative reasons. IER matters as democratic legitimacy of PA depends not only on legally enacted decisions, but on whether affected actors and oversight institutions can rebuild the grounds through which administrative authority is exercised (Lazar, 2024). In this vein, this contribution asks how the inferential turn disrupts reconstructable justificatory authority in public administration, and how institutional epistemic reconstructability can be conceptualized as a necessary condition for democratically justifiable administrative authority.

II) Epistemic power, traceability gaps, and institutional epistemic reconstructability

To explain how the inferential turn disrupts reconstructable justification in AI-mediated public administration, this contribution conceptualizes epistemic power as the operative mechanism through which inferential outputs acquire justificatory authority within administrative decision-making. Epistemic power is understood here as the double-ability to (i) influence what actors believe or know, and (ii) enable or disable epistemic influence within administrative contexts (Archer et al., 2020). Rather than focusing on epistemic domination or oppression (Dotson, 2014), this approach emphasizes how epistemic power structures the evidentiary, classificatory, and justificatory conditions under which administrative reasons become institutionally authoritative (Friberg-Fernros et al., 2022; Lazar, 2024). In this vein, the contribution distinguishes three interrelated epistemic power dimensions under AI-mediation: evidentiary epistemic power (what counts as relevant information), classificatory epistemic power (how subjects and cases become legible), and justificatory epistemic power (what counts as a valid administrative reason).

Analytically, the argument differentiates not only structural transformation (inferential turn) and operative mechanism (epistemic power), but also two distinct yet interrelated types of traceability gaps (TGs) as institutional failure modes along the justificatory chain linking inferential outputs to administrative authority: epistemic TGs arise when inferential transformations from data to model output cannot be meaningfully reconstructed; normative TGs, in turn, manifest when inferential outputs are reconstructable but cannot be translated into legally cognizable and democratically legitimate grounds for administrative action. Thus, unlike substantiation gaps (Verloop, 2026), TGs concern breakdowns in reconstructing how justification is produced, not merely what information informed a decision.

As such, TGs constitute institutional manifestations of insufficient institutional epistemic reconstructability – an institutional safeguard requiring (i) epistemic traceability (inferential pathways from data to output can be reconstructed in technically and institutionally intelligible terms), (ii) legal translation (outputs can be mapped onto reasons that count as valid grounds under the applicable legal framework), and (iii) adversarial contestability (those legally cognizable reasons are publicly defensible as legitimate exercises of administrative authority). Consequently, IER must not be confused with transparency, explainability, or auditability (Bücker et al., 2021; Loi et al., 2021): transparency may disclose model inputs or logic without showing how institutional actors translate outputs into legally valid reasons; explainability may render outputs intelligible without making authority-conferral processes contestable; and auditability may document compliance ex post without enabling affected actors to reconstruct the justificatory chain ex ante or challenge its normative adequacy.

Against this backdrop, this contribution argues that IER constitutes a necessary (albeit not sufficient) condition for democratic legitimacy under inferential governance. In this context, necessary means that democratically legitimate authority must remain answerable to those subject to it: without reconstructable reasons, affected stakeholders cannot assess whether administrative power was exercised in a way that respects public justification rather than mere statistical prediction. Importantly, this claim does by no means dis- or replace legality, accountability, or fairness; rather, it specifies the epistemic condition under which those criteria can be publicly (con)tested in AI-mediated administration.

The proposed framework can be further illustrated, for instance, in light of AI-assisted welfare-risk systems used for fraud investigations. These systems may satisfy transparency, explainability, and auditability requirements while still generating traceability gaps: epistemic TGs occur, for instance, when inferential transformations from administrative data to welfare-risk scores cannot be meaningfully retraced; normative TGs emerge, for example, when inferential outputs are technically reconstructable, yet institutions remain unable to justify why a specific probabilistic fraud score constitutes a legally sufficient reason for investigation or sanctions under welfare law. Thus, the key issue here is not opacity (Bignami, 2022; Bracci, 2023) alone, but inadequate institutional epistemic reconstructability. Table 1 summarizes the analytical framework of inferential governance in AI-mediated public administration, showing how epistemic power, TGs, and IER co-structure justificatory authority.

This contribution argues that preserving democratic legitimacy under the inferential turn requires institutional procedures capable of sustaining reconstructable justificatory authority. Corresponding procedural dimensions can be derived from IER core requirements: epistemic traceability procedures referring to the ability to retrace data sources and inferential operations; legal translation procedures pertaining to the capacity to map inferential outputs onto legally cognizable justificatory categories; and adversarial contestability procedures as the ability of affected actors, courts, or review bodies to challenge the epistemic-normative transformation underlying administrative decisions through procedural review, evidentiary challenge, or legal appeal. Taken together, these dimensions specify how reconstructable justification can be embedded within administrative practice.

Table 1: Inferential governance and the reconstruction of justificatory authority in AI-mediated PA

Analytical level	Core concept	Function in AI-mediated PA	Democratic implication
Structural transformation	Inferential turn	Administrative reasoning increasingly relies on probabilistic and classificatory inference generated within data infrastructures that partially operate independently from legally codified justificatory categories	PA justifications become less institutionally reconstructable
Operative mechanism	Epistemic power (evidentiary, classificatory, justificatory)	AI systems (re)shape the evidentiary relevance, classificatory legibility, and justificatory validity through which inferential outputs acquire administrative authority	The justificatory basis of PA authority becomes increasingly difficult to reconstruct, contest, and publicly legitimate
Institutional failure mode	Traceability gaps	Institutional breakdowns in reconstructing how inferential outputs acquire justificatory authority within administrative decision-making	Affected stakeholders cannot meaningfully reconstruct or contest administrative decisions
Institutional safeguard	Institutional epistemic reconstructability	Institutional capacity to retrace inferential operations, translate inferential outputs into legally cognizable justificatory categories, and contest authority-conferral processes	Necessary institutional condition for democratically justifiable administrative authority under inferential governance

Source: own compilation

III) Methodological framework of institutional epistemic reconstructability

Methodologically, the contribution adopts a conceptual and normative-institutional approach to analyse how AI-mediated inference transforms reconstructable justificatory authority within public administration. The analysis clarifies epistemic power, traceability gaps, and institutional epistemic reconstructability as interrelated dimensions within inferential governance while distinguishing IER from transparency, explainability, and auditability. On this basis, the contribution outlines a normative framework specifying the institutional conditions under which inferential outputs remain rebuildable as contestable administrative reasons within democratic governance. The paper's originality lies thus (i) in focusing on the institutional reconstructability of justificatory authority, and (ii) in introducing TGs and IER as a new way to capture conditions under which administrative reasons can remain democratically answerable.

IV) Four claims on inferential governance and democratic legitimacy

This contribution advances four analytically distinct yet interrelated core claims that map onto the different levels of the proposed analytical framework: structural transformation, operative mechanism, institutional failure mode, and institutional safeguard. Together, they specify how the inferential turn reconfigures administrative reasoning, promotes legitimacy-erosion, and points to IER as a necessary condition for democratic legitimacy.

1. *Inferential turn as structural transformation*. AI-mediated administration intensifies inferential mediation by shifting justificatory authority into probabilistic and classificatory infrastructures that operate at greater scale, speed, and institutional distance from legally codified justificatory categories. This reconfiguration renders administrative justifications less institutionally reconstructable, thereby posing a critical challenge to democratic legitimacy of PA.
2. *Epistemic power as operative mechanism*. AI systems reconfigure epistemic power along evidentiary, classificatory, and justificatory dimensions, thereby shaping what counts as relevant information, how subjects and cases become legible, and which considerations are recognized as valid administrative reasons.
3. *Traceability gaps as institutional failure mode*. Epistemic and normative traceability gaps emerge when institutions cannot reconstruct or translate how inferential outputs become justified grounds for administrative action. These gaps manifest as institutional failures to reconstitute the chain from data and model outputs to decisions, thereby undermining the ability of affected actors and oversight bodies to contest them.
4. *Institutional epistemic reconstructability as safeguard*. Democratic legitimacy under inferential governance depends on institutional conditions that enable epistemic traceability, legal translation, and adversarial contestability. IER thus represents a key safeguard, ensuring that inferential outputs can be reconstituted, contested, and justified as democratically legitimate reasons for administrative authority.

VI) Democratic legitimacy under institutional epistemic reconstructability

This contribution argues that democratic legitimacy under AI-mediated public administration significantly depends on the institutional conditions under which inferential outputs can be reconstructed as justificatory reasons for administrative action. In this vein, the analysis highlights IER as a key condition for democratically legitimate PA. What matters ultimately is not only whether AI systems can be inspected or explained, but whether the justificatory chain linking inference to administrative authority remains reconstructable within administrative practice. Without such reconstructability, PA risks becoming epistemically authoritative, thereby undermining its democratic grounding.

VI) References

Archer, A.; Cawston, A.; Matheson, B.; Geuskens, M. (2020): Celebrity, Democracy, and Epistemic Power, in: Perspectives on Politics, 18, 1, 27-42, DOI: https://doi.org/10.1017/S1537592719002615.

Beckman, L.; Hultin Rosenberg, J.; Jebari, K. (2024): Artificial intelligence and democratic legitimacy. The problem of publicity in public authority, in: AI & Society, 39, 975-984, DOI: https://doi.org/10.1007/s00146-022-01493-0.

Bignami, F. (2022): Artificial Intelligence Accountability of Public Administration, in: The American Journal of Comparative Law, 70, Issue Supplement_1, i312-i346, DOI: https://doi.org/10.1093/ajcl/avac012.

Binns, R. (2018): Algorithmic Accountability and Public Reason, in: Philosophy & Technology, 31, 543-556, DOI: https://doi.org/10.1007/s13347-017-0263-5.

Boughey, J. (2021): The Culture of Justification in Administrative Law: Rationales and Consequences, in: University of British Columbia Law Review 403, 54, 2, 4. SSRN, available at: https://ssrn.com/abstract=3954588 (accessed: May 13, 2026).

Bracci, E. (2023): The loopholes of algorithmic public services: an “intelligent” accountability research agenda, in: Accounting, Auditing & Accountability Journal, 36, 2, 739-763, DOI: https://doi.org/10.1108/AAAJ-06-2022-5856.
Bücker, M.; Szepannek, G.; Gosiewska, A.; Biecek, P. (2022): Transparency, auditability, and explainability of machine learning models in credit scoring, in: Journal of the Operational Research Society, 73, 1, 70-90, DOI: https://doi.org/10.1080/01605682.2021.1922098.
Burrell, J. (2016): How the machine ‘thinks’: Understanding opacity in machine learning algorithms, in: Big Data & Society, 3, 1, DOI: https://doi.org/10.1177/2053951715622512.
Cohen, M. (2010): The Rule of Law as the Rule of Reasons, in: Archiv für Rechts - und Sozialphilosophie, 96, 1, 1-16, available at: https://www.jstor.org/stable/23681150 (accessed: May 9, 2026).
Covilla, J. (2025): Artificial Intelligence and Administrative Discretion: Exploring Adaptations and Boundaries, in: European Journal of Risk Regulation, 16, 1, 36-50, DOI: https://doi.org/10.1017/err.2024.76.
Criado, J.; Valero, J.; Villodre, J. (2020): Algorithmic transparency and bureaucratic discretion: The case of SALER early warning system, in: Information Polity, 25, 4, 449-470, DOI: https://doi.org/10.3233/IP-200260.
Díaz de la Cruz, C.; Fernández Fernández, J.; Villegas-Galaviz, C. (2025): Model of ethical analysis of digital technologies: towards true digital humanism, in: AI & Society, 40, 5743-5754, DOI: https://doi.org/10.1007/s00146-025-02330-w
Dotson, K. (2014): Conceptualizing Epistemic Oppression, in: Social Epistemology 28, 2, 115-138, DOI: https://doi.org/10.1080/02691728.2013.782585.
Forst, R. (2012): The right to justification: elements of a constructivist theory of justice. New York: Columbia University Press.
Friberg-Fernros, H.; Karlsson Schaffer, J. (2024): An epistemic alternative to the public justification requirement, in: Philosophy & Social Criticism, 50, 6, 948-970, DOI: https://doi.org/10.1177/01914537221137855.
Lazar, S. (2024): Legitimacy, Authority, and Democratic Duties of Explanation, in: Sobel, D.; Wall, S. (eds.): Oxford Studies in Political Philosophy Volume 10, 28-56. Oxford: Oxford University Press, DOI: https://doi.org/10.1093/oso/9780198909460.003.0002.
Loi, M.; Ferrario, A; Viganò, E. (2021): Transparency as design publicity: explaining and justifying inscrutable algorithms, in: Ethics and Information Technology, 23, 253-263, DOI: https://doi.org/10.1007/s10676-020-09564-w.
Luise, G.; Cocozza, G. (2025): The process of “capacity building” in the public administration: artificial intelligence, transparency and new models of the administrative decision, in: Frontiers in Political Science, 7, 1605729, DOI: https://doi.org/10.3389/fpos.2025.1605729.
Oswald, M. (2018): Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power, in: Philosophical Transactions A: Mathematical, Physical and Engineering Sciences, 376, 20170359, DOI: https://doi.org/10.1098/rsta.2017.0359.
Simon, H. (1997): Administrative behavior: a study of decision-making processes in administrative organizations. New York: The Free Press.
Verloop, D. (2026): The Substantiation Gap. AI-Mediated Record Formation Under Administrative Law. SSRN, DOI: https://dx.doi.org/10.2139/ssrn.6164447.
Wihbey, J. (2025): AI and Epistemic Risk for Democracy: A Coming Crisis of Public Knowledge?. SSRN, DOI: https://dx.doi.org/10.2139/ssrn.4805026.
Zhu, X. (2026): Public administration with, of, and through AI: toward a new paradigm in the era of intelligence, in: Journal of Chinese Governance, 11, 1, 30-57, DOI: https://doi.org/10.1080/23812346.2025.2578589.

Statistical Lives, Identifiable Persons, and the Ethics of Predictive Policing

Korbinian Rüger | LMU Munich

Introduction

Predictive policing, i.e. the use of algorithmic systems to forecast criminal activity and identify individuals deemed at elevated risk, is typically justified on grounds of efficiency: more crimes detected, be>er resource allocation, more effective prevention. Modern AI capabilities only increase the pressure to rely on predictive methods in policing. Against these methods, critics have rightly raised concerns about algorithmic bias, racial discrimination, and violations of due process.1 This paper does not dispute these concerns. It aims, however, to identify a deeper structural problem that has received insufficient philosophical a>ention: even granting, for the sake of argument, that predictive policing increases the overall number of crimes detected or prevented2, it remains surprisingly difficult to say whether the transition to such a system is morally justified, and it is shown that this difficulty has a direct connection to the philosophical debate over identified versus statistical lives.

The Token-Shifting Problem

The efficiency argument for predictive policing typically operates at the level of types: more crimes of type X are detected, fewer crimes of type Y go unpunished, aggregate harm is reduced. But crimes, like harms more generally, do not occur at the level of types. They occur at the level of tokens, i.e. specific events involving specific individuals at specific times and places. Predictive policing does not merely change how many crimes are detected; it changes *which* crimes are detected. The algorithmic reallocation of police a>ention means that some crimes, some victims, some perpetrators, and some communities are drawn into the system that would not have been without algorithmic guidance. Simultaneously, other crimes that would have been discovered through conventional, less targeted policing remain undetected.

This token-shifting structure has a direct moral consequence: a system that appears superior at the aggregate level involves both winners and losers at the individual level. The victims whose cases are now detected benefit; but the victims whose cases are now missed, because police attention has been redirected elsewhere, are made worse off. The aggregate picture conceals a complex redistribution of both benefits and burdens across identifiable individuals.

It thereby fails the gold-standard of efficiency considerations, the Pareto criterion. A Pareto improvement is a change that makes someone be>er off witout making anyone worse off. A transition to predictive policing is not a Pareto improvement: there are losers, not merely winners. This by itself need not be a problem. There are very few policy changes that constitute Pareto improvements. However, it shows that the justification for switching to predictive policing is not as straightforward as some assume.
What is morally at stake, is whether the transition to predictive policing imposes losses on individuals that cannot straightfowardly be justified to those individuals by pointing to aggregate gains enjoyed by others. However, it could be the case that these individuals cannot be determined in advance or even after the fact.

Statistical vs. Identified Lives

This observation connects directly to the philosophical debate over identified and statistical lives. Following Schelling (1968), Fried (1969), and the more recent treatment by Cohen, Daniels, and Eyal (2015), we can distinguish between identified lives, i.e. concrete, particular individuals whose fates are at stake, and statistical lives, i.e. abstract, probabilistic beneficiaries or victims of policies assessed at the population level. It is well established that many people think that losses borne by “statistical individuals” carry less moral weight than losses borne by identified ones.

The moral significance of the identified/statistical distinction is itself contested, however.

There are strong reasons, both philosophical and intuitive, to think that the unidentifiability of losers does not diminish the moral weight of their losses. A policy that guarantees harm to some individual, without being able to say who, is not thereby rendered more justifiable than one that harms an identified person. What ma>ers morally is that real individuals bear real losses that cannot be justified to them, not that we can a>ach names to those individuals in advance.

This is relevant, because it is precisely the aggregate, statistical framing of the efficiency argument that obscures the individual losses at stake. The winners from the policy, however, are more easily identified after the fact. It is plausible, for example, that after successfully preventing a crime, police can say “this would not have been possible without predictive methods” and the specific potential victims are be>er off for it. It is much harder or maybe even impossible to identify the losers of the policy in a similar way.

The token-shifting problem, however, reveals that predictive policing's losers are not merely statistical abstractions. Rather, they are concrete individuals whose fates are determined by the policy, even if we cannot identify them. Rejecting the moral significance of the identified/statistical distinction then closes the escape route: the efficiency argument cannot rescue itself by pointing to the unidentifiability of the losers. Does this mean that the shift to predictive policing is unjustified?

Theory-Dependence

I argue that the answer is theory-dependent in an illuminating way. For a thoroughgoing utilitarian, for example, the token-shifting problem is not in itself a moral problem: what matters is aggregate welfare, and if predictive policing increases it, the redistribution of benefits and burdens across individuals is morally irrelevant. Similarly, ex ante contractualist views which evaluate policies by their expected benefits for each individual before outcomes are realized, may find predictive policing justifiable if it improves each person's ex ante prospects, regardless of who actually bears the costs ex post.

By contrast, views that are sensitive to how outcomes are distributed across individuals after uncertainty has been resolved, for example, ex post contractualist views, some prioritarian views, and certain deontological positions, will find the token-shifting problem morally significant. For such views, the fact that predictive policing guarantees losers, even if their identities are indeterminate at the time of decision, is not morally neutralized by aggregate gains. The losers have legitimate complaints that cannot be dismissed by pointing to benefits enjoyed by others.

The token-shifting structure of predictive policing thus functions as a discriminating test case for moral theory. Whether it constitutes a moral problem depends on prior commitments about the moral relevance of the identified/statistical distinction, the permissibility of imposing guaranteed losses on indeterminate individuals, and the weight one gives to individual complaints against aggregate welfare gains.

Implications

These considerations therefore do not straightforwardly establish that predictive policing is impermissible. They do establish that the standard efficiency argument is insufficient as a moral justification, and that the bar for justification is higher than proponents typically acknowledge. An adequate defense of predictive policing must grapple with the distribution of benefits and burdens across identified individuals, and with the moral significance of guaranteeing losers whose identities cannot be determined in advance. This paper aims to make that problem visible and to show that how seriously one takes it depends on deep commitments in moral theory that the debate over predictive policing has so far largely ignored.

Bibliography

Cohen, I. Glenn, Norman Daniels, and Nir Eyal (eds). 2015. “Statistical versus Identified Persons: An Introduction”, in I. Glenn Cohen, Norman Daniels, and Nir Eyal (eds), *Identified versus Statistical Lives: An Interdisciplinary Perspective*, Oxford University Press.

Fried, Charles. “Value of Life.” 1969. *Harvard Law Review 82* (7): 1415–37

Harcourt, Bernard E. 2005. “Against Prediction: Sentencing, Policing, and Punishing in an Actuarial Age”. *Public Law and Legal Theory Working Papers*, 2005.

Meijer, A., & Wessels, M. 2019. “Predictive Policing: Review of Benefits and Drawbacks”. *International Journal of Public Administration*, *42*(12), 1031–1039.

Purves, Duncan. 2022. “Fairness in Algorithmic Policing”. *Journal of the American Philosophical Association* 8, no. 4: 741–61.

Rüger, Korbinian. 2025. “Contractualism and Risk”. *Philosophy Compass* 20, no. 9.

Schelling, Thomas. 1968. “The Life You Save May Be Your Own.” In *Problems in Public Expenditure Analysis*, edited by Samuel B. Chase, 127–62. Washington, DC: Brookings Institution

SUB-SESSION

Gute wissenschaftliche Praxis als Governance-Problem: Wissenschaft und Lehre im Zeitalter generativer KI

20.11.2026, 09:00-10:30

Sprache: Deutsch

Chairs: Dirk Heckmann | bidt & TU München; Carolin Fuß | bidt

Generative KI verändert grundlegende Prinzipien guter wissenschaftlicher Praxis – insbesondere Autorschaft, Nachvollziehbarkeit und epistemische Verantwortung. Während KI-basierte Werkzeuge zunehmend selbstverständlich in Studium und Forschung eingesetzt werden, bleiben institutionelle und normative Leitplanken oft unklar oder hinter der technologischen Entwicklung zurück.
Die Sub-Session adressiert diese Spannung als interdisziplinäres Governance-Problem und fragt, wie gute wissenschaftliche Praxis unter Bedingungen allgegenwärtiger KI neu bestimmt, vermittelt und institutionell abgesichert werden kann.

Das 90-minütige Diskussionspanel soll VertreterInnen unterschiedlicher Fachrichtungen zur Frage des KI-Einsatzes in Lehre, Studium, Wissenschaft und Forschung zu Wort kommen lassen. Insbesondere die Frage der guten wissenschaftlichen Praxis und ihrer Vermittlung bereits im Studium soll Raum eingeräumt werden.

Panel-Diskussion:

- Hjördis Czesnick | OWID
- Marion Händel | Hochschule Ansbach
- Dirk Heckmann | bidt & TU München
- Lena Kästner | Universität Bayreuth

SUB-SESSION

Zwischen Code und KI: Wandel der Arbeit in der Softwareentwicklung

20.11.2026, 12:30-14:00

Sprache: Deutsch

Chairs: Angela Graf | bidt; Florian Butollo | Goethe-Universität Frankfurt; Marco Blank | FAU Erlangen-Nürnberg

Die Softwareentwicklung gilt als ein zentrales Anwendungsfeld und zugleich als Vorreiter bei der Nutzung von Künstlicher Intelligenz (KI). Der zunehmende Einsatz von KI-Tools verändert dabei gegenwärtig sowohl die technischen als auch die sozialen Dimensionen dieses Berufs- und Tätigkeitsfelds. Dies wirft grundlegende Fragen zur Zukunft von Arbeit, Qualifikation und professioneller Identität von Softwareentwickler:innen auf.
Die Sub-Session widmet sich diesen Transformationsprozessen aus einer interdisziplinären Perspektive und adressiert insbesondere folgende Fragestellungen:

- Welche neuen Entwicklungen zeichnen sich im Berufsfeld der Softwareentwicklung durch den Einsatz von KI ab?
- Wie verändern sich Arbeitspraktiken von Softwareentwickler:innen und Arbeitsprozesse in der Softwareentwicklung?
- Inwiefern beeinflusst KI die Effizienz, Performance und Effektivität von Entwicklungsarbeit?
- Welche Auswirkungen hat der Einsatz von KI-Tools auf die Arbeitsqualität und wie wird damit umgegangen?
- Wie verändern sich Kompetenzprofile, Qualifikationsanforderungen und das professionelle Selbstverständnis von Softwareentwickler:innen?

Die Sub-Session bringt Beiträge aus der Arbeits-, Technik- und Digitalisierungsforschung zusammen, die sich mit den Auswirkungen von KI auf Softwareentwicklungsarbeit befassen. Die einzelnen Beiträge beleuchten unterschiedliche Facetten des Themenfelds. Ziel der Sub-Session ist es, diese Entwicklungen theoretisch fundiert und empirisch gestützt zu reflektieren und ihre Implikationen für das Berufsfeld der Softwareentwicklung zu diskutieren.

Vorträge:

Die Automatisierung der Automatisierung? Generative KI und die Refiguration von Programmierarbeit

Florian Butollo | Goethe-Universität Frankfurt; Ann Katzinski | WZB & Weizenbaum-Institut; Marlene Kulla | WZB/Weizenbaum-Institut

Der Beitrag untersucht, wie generative KI Arbeitspraktiken, Qualifikationsanforderungen und das berufliche Selbstverständnis in der Softwareentwicklung verändert. Grundlage sind 10 Expert:inneninterviews, 4 Fallstudien sowie eine netnographische Analyse globaler Reddit-Beiträge von Entwickler:innen.
Tools wie Cursor oder GitHub Copilot automatisieren routinisierte Tätigkeiten wie die Erstellung von Boilerplate-Code oder Dokumentation, ohne menschliche Arbeit jedoch schlicht zu ersetzen. Vielmehr verschiebt sich der Schwerpunkt der Arbeit auf wiederkehrende Toolanpassung, Prompting, Evaluation und Qualitätssicherung. In unserer Untersuchung zeigt sich dabei eine Verlagerung vom eigentlichen Schreiben von Code hin zur Beaufsichtigung, Einordnung und Orchestrierung maschinell generierter Vorschläge.
Zugleich erweitert generative KI technische Handlungsmöglichkeiten, etwa durch schnellere Iterationen und erleichtertes Experimentieren, erzeugt aber auch neue Anforderungen. Hierzu zählen kognitive Zusatzarbeit, interpretative Kompetenzen sowie ein praktisches Verständnis der Funktionsweise und Grenzen von Modellen. Die Integration generativer KI verändert damit nicht nur Arbeitspraktiken, sondern auch berufliche Rollenbilder

und Erwartungsstrukturen, insbesondere mit Blick auf Junior-Beschäftigte und Neueinstellungen. Damit verbinden sich Sorgen um Dequalifizierung, verschobene Lernprozesse und eine mögliche Intensivierung von Arbeit. Der Beitrag argumentiert, dass generative KI Programmierarbeit nicht obsolet macht, sondern sie in Form hybrider Mensch-Maschine-Interaktionen neu organisiert. Die Ergebnisse sprechen somit gegen einfache Substitutionsnarrative und plädieren für ein differenziertes Verständnis der „automatisierten Automatisierung" als vielschichtige Rekonfiguration digitaler Facharbeit.

Frisst die KI ihre Erschaffer? Was KI mit der Arbeit in Software-Entwicklung und Programmierung macht

Sabine Pfeiffer | bidt & FAU Erlangen-Nürnberg; Marco Blank | FAU Erlangen-Nürnberg; Angela Graf | bidt

Auf Basis einer standardisierten Befragung unter Arbeitnehmer:innen im Bereich IT und Engineering (mit mindestens 20% ihrer Arbeitszeit in Programmier- und Softwareentwicklungstätigkeiten) bietet der Beitrag einen systematischen Überblick über die Verbreitung, Nutzung und (wahrgenommene) Auswirkungen KI-gestützter Tools in der Softwareentwicklung. Dabei wird erstens aufgezeigt, in welchem Umfang unterschiedliche KI-Tools in der Praxis zum Einsatz kommen sowie in welcher Art und Weise und für welche konkreten (Teil-)Aufgaben sie im Softwareentwicklungsprozess genutzt werden. Zweitens untersucht der Beitrag die Auswirkungen des KI-Einsatzes auf Arbeitsbelastung und -entlastung sowie auf wahrgenommene und tatsächliche Produktivität und Effizienz. Drittens werden die Auswirkungen auf das berufliche Selbstverständnis sowie auf die Qualifikationsanforderungen beleuchtet.

Der Beitrag zeichnet damit ein detailliertes Bild der Arbeitsteilung zwischen Mensch und KI im Kontext der Softwareentwicklung, gibt Aufschluss über Veränderungen von Arbeitsprozessen und bietet damit eine breite empirische Grundlage für ein vertieftes und differenziertes Verständnis der Transformation von Softwareentwicklungsarbeit durch KI.

KI, Automatisierung und Arbeitsqualität in der Softwareentwicklung

Martin Krzywdzinski | WZB/Weizenbaum-Institut/Helmut-Schmidt-Universität Hamburg

Der Einsatz von KI beschleunigt Automatisierungstendenzen in der Softwareentwicklung. Wie schätzen Softwareentwickler selbst die Automatisierungsdynamik ein und wie bewerten sie die Auswirkungen auf ihre Arbeitsbedingungen? Diese beiden Fragen werden basierend auf einer 2026 durchgeführten internationalen Befragung von 1.731 Softwareentwicklern analysiert. Die Analyse diskutiert die Entwicklungen im Hinblick auf Unterschiede zwischen verschiedenen Berufsfeldern in der Softwareentwicklung und verschiedene Tätigkeitstypen.

SUB-SESSION

Generative KI als Interventionistin und Sparringspartnerin: Interaktionsdynamiken in digitalen Öffentlichkeiten verstehen und stabilisieren

20.11.2026, 12:30-14:00

Sprache: Deutsch

Chairs: Steliyana Doseva | bidt; Sofie Fleischer | bidt; Hannah Schmid-Petri | Universität Passau & bidt; Andreas Wenninger | bidt; Katharina V. Hajek | LMU München; Carsten Reinemann | LMU München

Digitale Öffentlichkeiten und ihre spezifischen Funktionslogiken haben in den letzten Jahren zu einem Wandel öffentlicher (online-)Diskurse über umkämpfte Themen geführt. Anonymität, Emotionalisierung und algorithmische Inhalteselektion begünstigen dort Dynamiken (affektiver) Polarisierung (Iyengar et al., 2019), etwa Lagerbildung und Identitätskonflikte (Esau, 2025). Dadurch besteht die Gefahr, dass sich in der Folge insbesondere moderate Stimmen aus politischen Diskursen zurückziehen (Mutz, 2006; Brennauer et al., 2024). Gleichzeitig zeigt sich eine wachsende Feindlichkeit gegenüber journalistischen Medien und ihren Akteur:innen, die Informationen bereitstellen, einordnen und moderieren. Beide Entwicklungen gefährden die deliberative Grundlage demokratischer Gesellschaften, die öffentliche Verhandlung verschiedener Positionen brauchen (Lutrell et al., 2019).

Large Language Models (LLMs) und darauf basierende Anwendungen eröffnen neue Möglichkeiten, solche Dynamiken nicht nur zu analysieren, sondern in sie einzugreifen oder sie durch Simulation kontrolliert erfahrbar zu machen (Jungherr & Rauchfleisch, 2025). In zwei Teilen wird in der Sub-Session diskutiert, inwiefern generative KI dazu beitragen kann, Interaktionsdynamiken in digitalen Öffentlichkeiten zu verstehen und die deliberative Qualität von Online-Diskursen zu unterstützen.

Im ersten Teil beleuchtet Prof. Dr. Andreas Jungherr, wie generative KI die Informationsvermittlung und politische Kommunikation verändert und neue Fragen nach Machtverhältnissen und Handlungsmöglichkeiten in digitalen Öffentlichkeiten aufwirft. Daran anknüpfend wird der Einsatz generativer KI im Kontext der bidt-Projekte CRASH und DemocraGPT diskutiert. Dabei werden zwei konkrete Rollen generativer KI gegenübergestellt: CRASH untersucht KI als **Interventionistin**, die als dialogbasierte KI-Anwendung medienfeindlichen Überzeugungen begegnet und versucht, diese abzuschwächen. DemocraGPT untersucht generative KI als **Sparringspartnerin** für schwierige politische Gespräche, um theoretisch relevante Konflikt- und Widerstandsdynamiken kontrolliert erfahr- und reflektierbar zu machen.

Im zweiten Teil werden auf dieser Grundlage in einer offenen Fishbowl-Diskussion Erfolgsbedingungen und Grenzen solcher KI-Anwendungen, mögliche methodische Designs sowie ethische Herausforderungen erörtert. Prof. Dr. Andreas Jungherr diskutiert gemeinsam mit Dr. Dennis Frieß, Experte für Online-Deliberation und digitale Partizipation, und Dr. Ursula Alexandra Ohliger, die Perspektiven aus politischer Kommunikationsforschung und KI-Transfer einbringt. Die Fishbowl ist ausdrücklich als Diskussionsformat mit dem Publikum konzipiert: Die Konferenzteilnehmenden sind eingeladen, freie Plätze im inneren Kreis einzunehmen, Fragen einzubringen und Annahmen der Projekte kritisch zu diskutieren.

Impulsvorträge & Fishbowl-Diskussion:

Generative KI und die Neuordnung digitaler Öffentlichkeit

Andreas Jungherr | bidt & Universität Bamberg

Generative KI entwickelt sich zu einem wichtigen Element digitaler Öffentlichkeit und verändert die Handlungsmöglichkeiten der an ihr beteiligten Akteure. Dienste auf Basis generativer KI werden zunehmend zu einem wichtigen Zugang zu Informationen und zu einer alltäglichen Anlaufstelle für Fragen über die Welt. Damit entwickeln sich KI-Anbieter zu neuen politischen Gatekeepern, die in Konkurrenz mit digitalen Plattformunternehmen treten und Informationen nicht nur selektieren, sondern auch aufbereiten und interpretieren. Zugleich wird generative KI gezielt von politischen Akteuren eingesetzt, um mit Menschen in den Austausch zu treten und ihre Meinungen und ihr Verhalten zu beeinflussen. Gleichzeitig eröffnet generative KI Menschen neue Möglichkeiten, ihre Informationsumgebungen und -flüsse aktiv und selbstbestimmt zu gestalten. Damit erfüllt sich – zumindest für einige – das Informationsversprechen der Digitalisierung in bislang nicht gekanntem Maße. Diese Entwicklungen werfen grundlegende Fragen nach der Verteilung von Macht und Handlungsmöglichkeiten in digitalen Öffentlichkeiten auf. Für die Sozialwissenschaften gilt es, diese Transformationen frühzeitig zu verstehen, ihre gesellschaftlichen und politischen Konsequenzen zu identifizieren und zu ihrer Gestaltung beizutragen.

KI als Trainerin: Konflikthafte Gespräche simulieren und reflektieren

Katharina V. Hajek | LMU München

Politische Gespräche sind ein wichtiger Bestandteil demokratischen Miteinanders: Menschen, die sich über unterschiedliche Perspektiven austauschen stärken ihre Urteilsbildung, beteiligen sich häufiger und erleben eine höhere politische Wirksamkeit (Landemore, 2024; Schmitt-Beck & Schnaudt, 2023). Gleichzeitig zeigen Studien dass gerade kontroverse politische Gespräche oft mit sozialen und psychologischen Kosten verbunden und deshalb vermieden werden können (Iyengar et al., 2012; McLaughlin et al., 2022; Santoro & Broockman, 2022). Das Projekt DemocraGPT hat deswegen das Ziel, mit generativer KI einen kontrollierbaren Erfahrungsraum für schwierige politische Gespräche bereitzustellen, dessen Nutzung die Bereitschaft zum Austausch mit Andersdenkenden stärken kann. Grundlage dafür sind Studien zum Einsatz von LLMs für demokratische Deliberation, Förderung von Common Ground und positiven Intergruppenwahrnehmungen (Argyle et al., 2023; Goldberg et al., 2024; Hruschka & Appel, 2026), sowie etablierte Ansätze erfahrungs- und reflexionsbasierten Lernens (Boud et al., 2013; Dewey, 1986; Kang & Lakshmanan, 2026). Auf der bidt-Konferenz stellt DemocraGPT zwei zentrale Bestandteile vor: Erstens zeigt das Forschungsteam am aktuellen DemocraGPT-Prototyp, wie sozialpsychologische Theorie systematisch in dynamisches LLM-Verhalten übersetzt werden kann. Zweitens adressiert das Projekt die weiterführende theoretisch-technologische Frage der Übertragbarkeit von Mensch-KI-Erfahrungen auf spätere Mensch-Mensch-Kommunikation. DemocraGPT entwickelt hierfür ein mehrdimensionales Transferverständnis, das emotionalen, kognitiven, motivationalen, kommunikativen und demokratischen Transfer sowie AI-spezifische Boundary Conditions unterscheidet.

KI als Interventionistin: KI-gestützte Interventionen in digitalen Öffentlichkeiten

Steliyana Doseva | bidt

Krisen wie die COVID-19-Pandemie oder der Klimawandel verdeutlichen, welche gesellschaftlichen Folgen die Verbreitung von Falschinformationen haben kann. Neben Mis- und Desinformation über Ursachen, Folgen und den angemessenen Umgang mit Krisen sind digitale Öffentlichkeiten dabei auch von Narrativen geprägt, die sich gegen etablierte journalistische Medien richten, ihnen pauschal unterstellen manipulativ und als Werkzeuge der „Mächtigen" zu agieren und deren Glaubwürdigkeit grundsätzlich infrage stellen. Solche pauschalen Vorwürfe werden in der Forschung unter dem Begriff der Medienfeindlichkeit bzw. des Medienzynismus gefasst (Schindler et al., 2018; Schultz et al., 2021).

Der Beitrag setzt an der Frage an, wie medienfeindlichen Überzeugungen und Aussagen im digitalen Raum begegnet werden kann und welches Potenzial dabei dialogbasierte KI-Anwendungen bieten. Einen Ausgangspunkt bilden aktuelle Befunde aus der Desinformations- und Verschwörungsforschung: So zeigen Costello et al. (2024), dass individuell geführte Dialoge mit einem Sprachmodell verschwörungstheoretische Überzeugungen nachhaltig abschwächen können. Dabei könnte das Potenzial dialogbasierter Interventionen über die Vermittlung von Gegenargumenten hinausgehen: Die Überschätzung des eigenen Verständnisses komplexer Zusammenhänge steht mit Verschwörungsüberzeugungen in Verbindung (Vitriol & Marsh, 2018), während die Anregung zur Reflexion über die Begründung eigener Überzeugungen einen Ansatzpunkt für deren Veränderung bieten kann (Meyer et al., 2024). Daran anknüpfend wird diskutiert, inwiefern sich dialogbasierte KI auch als Intervention gegen Medienfeindlichkeit einsetzen lässt. Im Mittelpunkt stehen die Fragen, wie eine solche KI auf medienfeindliche Narrative reagieren sollte, welche Mechanismen dialogischer Intervention erfolgversprechend sein könnten und wie sich deren Wirkung empirisch untersuchen lässt. Damit entwickelt der Beitrag konzeptionelle und methodische Perspektiven für den Einsatz von KI als Interventionistin in medienfeindlich geprägten digitalen Diskursen.

Fishbowl-Diskussion:

- Andreas Jungherr | bidt & Universität Bamberg
- Dennis Frieß | Heine Center for Resilient Democracy (HeiCReD), Heinrich-Heine-Universität Düsseldorf
- Ursula Alexandra Ohliger | Deutsches Forschungszentrum für Künstliche Intelligenz

SUB-SESSION
Interdisciplinarity for the Study of Socially Relevant Imagery

20.11.2026, 12:30-14:00
Language: English
Chair: Mario Haim | LMU Munich

The modern digital environment is increasingly characterized by the prevalence of visual content, such as memes, videos, and news imagery. Being widely distributed, visual content plays an ever more significant role in shaping public discourse and influencing human opinion across domains, ranging from environmental communication to electoral processes. Together with the large-scale availability of such data and the increase in computational resources, these developments underscore the importance of extending recent practices in digital transformation research to visual content, thereby requiring a highly interdisciplinary approach that integrates expertise from both the social and information sciences. The complexities involved in integrating disciplines within this emerging field raise the question: *What role does interdisciplinarity play in digital transformation research, and how can it be translated into research practice?*
To address this question, we examine the role of interdisciplinarity in digital transformation research, focusing on two projects around socially relevant imagery: KLIMA-MEMES and ToxicAInment. The first part of the panel features presentations from these projects showcasing results as well as challenges from these interdisciplinary collaborations. These presentations include perspectives from social and computer sciences, focusing on studying visual data from social media to answer research questions of societal relevance. In the second part of the sub-session, these perspectives are then integrated in a panel discussion where representatives from both projects and invited experts will discuss the challenges for as well as the benefits of successful interdisciplinary research in the field of digital transformation.

Presentations:

What Media Frames Reveal About Stance: A Dataset and Study of Memes in Climate Change Discourse
Shijia Zhou | Center for Information and Language Processing (CIS), LMU Munich

Media framing refers to the emphasis on specific aspects of perceived reality to shape how an issue is defined and understood, often in alignment with authors' opinions and stances. This study applies an interdisciplinary approach to conceptualize and computationally investigate the interaction between stance and media frames in internet memes on climate change. We introduce *CLIMATEMEMES*, the first dataset of climate change memes (N = 1,184) annotated with both stance and media frames, drawing on research in communication science. We propose two meme understanding tasks: stance detection and media frame detection; and evaluate vision-language models (VLMs), including LLaVA-NeXT and Molmo. Our findings show that VLMs perform well on stance detection but struggle with media frame identification. Finally, we analyze the limitations of VLMs in capturing nuanced expressions of stance and framing in climate change-related internet memes.

Conceptualizing and Detecting Toxic Entertainment in Visual Social Media Content
Ursula K. Schmid | LMU Munich; Philipp Spielbauer | LMU Munich

Intolerance is increasingly embedded in visual entertainment formats that can obscure its harmful nature and complicate detection. This is especially true on multimodal platforms like TikTok, where toxic messages are layered across a variety of visuals (e.g., symbols, stickers, remixed clips), accompanied by audio and text. Often, messages rely on their combination of visuals and text to convey the toxic meaning that no single

modality would reveal alone. This presentation tackles the challenge of conceptualizing and automatically detecting such toxic entertainment in visual social media content. Drawing from previous findings of a qualitative multimodal content analysis of TikToks posted during the 2024 Saxony state election, we developed a typology of toxic entertainment, informing a codebook to detect content at the intersection of violence and entertainment in visual social media posts. We then applied the codebook to TikTok data collected via a keyword-based search for the German national election 2024. Finally, we present initial results on automating the detection of two subcategories central to toxic entertainment, namely violence and humor, using VLMs within the election dataset.

Where Have All the Memes Gone? Analyzing the Prevalence and Virality of Memes Using Vision-Language Models

Nadezhda Ozornina | Department of Media and Communication, LMU Munich

This study evaluates core, largely untested views on the prevalence, antagonistic nature, and virality of memes in social media environments. We first benchmark five open-access VLMs for large-scale meme detection and classification, where Qwen 2.5-VL-7B-Instruct proves most effective and was deployed to analyze climate-related posts on Facebook (*N* = 726,557) during the 2023 and 2024 UN Conferences of the Parties (COP). Based on the annotations, we offer preliminary insights for testing theoretical assumptions about memetic content. First, memes are not ubiquitous, accounting for less than 4.5% of image-based posts. Second, memes are more viral than other visual posts, suggesting their efficiency as a mobilization multiplier. Third, the characterization of memes as pervasive “discursive weapons” is questioned by the marginal prevalence of offensive memes (<1%) and their lower share rate compared to non-offensive memes. These findings call for a cautious reassessment of claims about meme ubiquity, providing insights into the capabilities of VLMs for analyzing memetic content.

Panel Discussion:

- Willem Fourie | School for Data Science and Computational Thinking, Stellenbosch University, South Africa
- Diana Rieger | Department of Media and Communication, LMU Munich
- Johannes Schusterbauer | Department of Computer Science, LMU Munich

SUB-SESSION

Digital Competences: Insights and Strategic Implications for the Empowerment of Society for the Digital Future

20.11.2026, 14:00-15:30

Language: English

Chairs: Roland A. Stürz | bidt; Danilo Harles | bidt

Digital competences are an elementary prerequisite for sustainable participation in private and professional life within a society undergoing continuous and profound transformation due to digitalization. Even at a young age, children require digital skills to navigate the opportunities and risks associated with digital technologies (Zimmermann & Tomczyk, 2025). In professional life, the digital transformation has resulted in fundamental structural changes and a continuous evolution of competency requirements for employees. Consequently, existing professional skills often decline in significance, while new – specifically digital – competences gain increasing relevance. It is estimated that in almost 90% of all jobs some digital device or machine is used and therefore at least some degree of digital skills is required (Cedefop, 2022). Currently, 91% of the workforce in Germany considers digital competences to be important or very important for their professional lives (Stürz et al., 2025). Furthermore, digital skills are becoming increasingly vital after retirement to ensure access to digital services and continued participation in social life. Additionally, digitalization fosters innovations – such as in the healthcare sector – that address the specific needs of older people. However, a sufficient level of digital literacy remains a prerequisite for utilizing these innovations (Expertenkommission Forschung und Innovation, 2023).

Accordingly, there are numerous initiatives and measures at the regional, national, and European levels aimed at strengthening digital skills. Of particular note is the **Digital Decade Strategy** (2021–2030) at the EU level, which aims to ensure that at least 80% of adults in Europe possess basic digital skills by 2030 (European Parliament, 2022). Other measures on a European level include the **Union of Skills** as an overarching strategy to enhance education, training, and workforce skills, including digital competences (European Commission, 2026), or the **AI Continent Action Plan** which aims to reinforce AI skills (European Commission, 2025). Additionally, various EU member states, including Germany, have developed different digitalization strategies with digital skills often being a key pillar of those strategies (European Union, 2026). Other significant measures include the **Bavarian Alliance for Digital Competencies**, a partnership between the Ministry for Digital Affairs in Bavaria and various companies providing free, high-quality educational resources (Bayerisches Staatsministerium für Digitales, 2025), and initiatives such as **fit4internet** in Austria, which promotes the competent use of digital technologies and broad societal participation in digitalization (fit4internet, 2026a).

A fundamental requirement for the design and evaluation of measures to expand digital competences is the accurate assessment of such abilities. From a scientific perspective, the primary questions concern how digital competences can be defined and subsequently measured. Once assessed, the focus shifts to what can be learned from the measurement and distribution of these competences and what implications can be derived for improvement measures.

In scientific research, various definitions of digital competences and approaches to measuring them exist. The sheer variety of different terms used – such as "digital literacy," "digital competence," "e-skills," or "ICT skills" – for the same construct, or conversely, different constructs behind the same term, complicates a uniform definition and, consequently, its operationalization (Ferrari et al., 2012; Oberländer et al., 2020). One of the most widely used reference models for defining and assessing digital competences in research (Oberländer et al., 2020), but also in policy and practice (Centeno & Cosgrove, 2025) is the European Digital Competence Framework (DigComp), which has been updated several times since its initial publication in 2013. Within scientific research, DigComp is adapted and operationalized differently depending on the research objective

and context. This variation reflects fundamental tensions of digital competences assessment along several dimensions, including the balance between standardization and contextual validity, between measurement efficiency and comprehensiveness, and between feasibility and validity (e.g. between various possible forms of competence measurement, such as self-reported and performance-based assessments (Beimborn & Hildebrandt 2022).

The contributions to the subsession **"Digital Competences: Insights and Strategic Implications for the Empowerment of Society for the Digital Future"** will discuss and reflect on how the different DigComp-based approaches handle these tensions, and how the application of the framework shapes how digital competences are understood, compared and interpreted in research, policy and practice. In their presentations and the panel talk, the speakers will bring together theoretical, methodological and policy-oriented perspectives on the assessment of digital competences in order to examine how the concept can be defined and validly measured as well as how the resulting findings can inform efforts to strengthen digital competences across society and reduce associated digital inequalities.

Presentations & Panel Discussion:

The DigComp 3.0 update: Why, How, and the Framework's Potential for Defining and Assessing Digital Skills

Jude Cosgrove | Joint Research Centre, European Commission

Significant digital technological developments and concerns have emerged since the publication of the fourth edition of DigComp 2.2 in 2022 (Vuorikari et al., 2022a). This presentation describes the context, rationale and methodology for updating the framework to the fifth edition, DigComp 3.0 (Cosgrove & Cachia, 2025). The policy context for the update, which is grounded in the European Declaration on Digital Rights and Principles (European Parliament, 2023), is described. The presentation will explain why and how DigComp 3.0 has been updated including five key priority themes (AI; mis- and disinformation; cybersecurity; rights, choice and responsibilities; and wellbeing in digital environments), novel features, and the multi-method approach used. It will then move to a focus on digital skills assessment, outlining some of the key principles of assessment (Newton, 2007; Stobart, 2008; Cook & Pitoniak, 2025; William, 2006; AERA, APA & NCME, 2014; Black & Wiliam, 1989; Sadler, 1989, 2010) and a brief overview of existing assessments which have been developed on the basis of DigComp (e.g. Clifford et al., 2020; Mattar et al., 2022; Vuorikari et al., 2022b). The presentation concludes with a critical commentary on the assessments' purposes, alignment to the identified principles, and implications for future work in this area.

Measuring and Status Quo of Digital Literacy Among Older Adults

Linda Göbl | Catholic University of Applied Sciences Freiburg; Michael Doh | Catholic University of Applied Sciences Freiburg

Successful and independent ageing in an increasingly digital world requires digital literacy. Media studies has identified various forms of digital inequality between younger and older adults, as well as intersectional ones within older age groups. Measuring digital literacy is a prerequisite for addressing competence-based digital inequalities in later life and for developing targeted counseling, educational, and support services. However, which dimensions and areas of digital literacy are relevant to the lifeworld of older adults? And how can digital literacy be validly measured in the context of heterogeneous digital ageing, especially given that access to ICT, usage opportunities and skill requirements in the digital world are constantly evolving? A research group at the Catholic University of Applied Sciences Freiburg has developed a quantitative measurement instrument specifically applicable to older adults while remaining comparable to younger age groups. The foundation was

an 82-item self-assessment tool developed by the European Commission in 2020: DigCompSAT (Clifford et al., 2020). Through modeling and validation, a methodologically robust yet practical short version consisting of 14 items was developed. This version was implemented in Germany in 2024 as part of a representative study of individuals aged 60 and above (Senior – Information – Media, SIM 2024, n = 2,000). The presentation highlights age-specific aspects of digital literacy and describes the development of the modified measurement instrument for assessing digital literacy in later life. It includes selected findings from the SIM 2024 study (Rathgeb et al., 2025), with particular attention to the oldest age groups. Beyond the empirical results, the presentation will stimulate discussion on conceptual perspectives regarding the assessment and operationalization of digital literacy in later life. In particular, it invites reflection on the tension between standardized measurement and age-sensitive adaptation, between self-assessed and performance-based competences, and between competence models developed for the general population and the heterogeneous realities of ageing. The presentation concludes with implications for educational research and practice.

Digital Competences Around the Globe: A Comparative Study of Nine Countries

Roland A. Stürz | bidt; Danilo Harles | bidt

The process of digital transformation does not stop at national borders but is inherently transnational. In this context, the digital competence level of the population and the workforce in a country constitutes a key resource for navigating the digital transformation and ensuring economic prosperity. At the same time, digital competences are unevenly distributed. Within countries, such inequalities can hinder individual participation and reduce innovative capacity and economic competitiveness. Assessing the distribution of digital competences within and between countries is therefore crucial for understanding how Germany is positioned in the digital transformation. The presentation first focuses on challenges of operationalizing digital competences, especially with regard to adapting existing instruments to current developments in digitalization, and subsequently assessing them in a research-economical, efficient manner. Against this background digital competences in this study are measured using a shortened version of the DigComp self-assessment tool (DigCompSAT Mini; Vuorikari et al., 2026). In addition, an AI competence scale developed by the bidt in alignment with the DigCompSAT is used. Key findings of the analyses of representative data collected by the bidt in nine countries using online surveys (Germany, France, Italy, Spain, Finland, UK, USA, China, Taiwan; each country with n = 1,000) will be presented. The presentation provides a novel, globally comparative perspective on digital competence research. It offers a unique empirical basis for large-scale benchmarking of digital competences and for assessing associated differences within and between various societies, thereby revealing broad structural patterns of digital inequality.

From Frameworks to Action: Operationalising and Measuring Digital Competence for Skills Policy and Workforce Development

Ulrike Domany-Funtan | Secretary General fit4internet, Austria

While digital competence frameworks such as DigComp have become an important reference point for research, education and policy, significant challenges remain regarding their conceptualization, operationalization and measurement. Different assessment approaches focus on self-assessment, knowledge, attitudes, task performance or behavioral indicators, each capturing different dimensions of digital competence and potentially leading to different conclusions about the actual digital readiness of individuals, workforces and societies. Drawing on the Austrian Digital Skills Barometer 2025/2026 (n = 2,273; fit4internet, 2026b) and fit4internet's experience in designing and implementing national digital skills initiatives, this presentation explores how digital competence frameworks can be translated into measurable constructs that are meaningful not only for research purposes but also for skills governance, workforce development and evidence-based policy making. Particular attention will be paid to the relationship between digital activities, digital knowledge and digital attitudes as well as to the opportunities and limitations of different measurement approaches. The

presentation will discuss how these methodological choices influence the interpretation of digital competence levels and their implications for labor markets, organizational transformation and national competitiveness. In addition, the presentation will introduce first results from a newly available cross-national comparison between Austria and Germany, based on data from the Austrian Digital Skills Barometer and an analogous survey conducted in Germany (n = 4,000). The presentation analyses similarities and differences in digital competence profiles, attitudes towards digital transformation, emerging technology-related competences, digital resilience, and learning behavior. Complementing broader cross-national analyses, this approach focuses on a small number of countries, allowing for a stronger emphasis on the macro-level context of digital competence distributions. This perspective provides valuable insights into how structural differences shape digital competences and helps identify starting points for targeted policies to reduce competence-based digital inequalities.

Panel Discussion: Empowering Society for the Digital Age: From Measuring Competences to Strategic Action

The session will conclude with a panel discussion among presenters and the audience focusing on advantages and disadvantages of DigComp as a framework to define digital competences, the differing methodological approaches to their operationalization and measurement in the presented studies, and the research findings and their overarching implications for how society can be effectively supported and prepared for the future in terms of digital competences.

References

AERA, APA & NCME (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education) (2014). Standards for educational and psychological testing. Washington DC. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf [14.04.2026].

Bayerisches Staatsministerium für Digitales (2025). Allianz Digitale Kompetenzen. https://www.stmd.bayern.de/themen/allianz-digitale-kompetenzen/ [14.04.2026].

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. In: Assessment in Education: Principles, Policy & Practice, 5(1), 7–74. https://doi.org/10.1080/0969595980050102.

Beimborn, D./Hildebrandt, Y. (2022) Scientific Approaches to Identify and Measure Digital Competences. In Ramin, P. (Ed.) Digital Competence and Future Skills. Munich, 3-24. https://doi.org/10.3139/9783446474284.001.

Cedefop (2022). Setting Europe on Course for a Human Digital Transition. Luxembourg. https://doi.org/10.2801/253954 [14.04.2026].

Centeno, C./Cosgrove, J. (2025). Ten Years of DigComp: A Framework More Essential Than Ever. Ispra. https://publications.jrc.ec.europa.eu/repository/handle/JRC143430 [14.04.2026].

Clifford, I. et al. (2020). DigCompSat. Luxembourg. https://doi.org/10.2760/77437 [14.04.2026].

Cook, L.L./Pitoniak, M.J. (2025). Educational measurement. (5th ed.). Oxford. http://dx.doi.org/10.1093/oso/9780197654965.001.0001.

Cosgrove, J./Cachia, R. (2025). DigComp 3.0: European digital competence framework. Luxembourg. [14.04.2026]. https://doi.org/10.2760/0001149.

European Commission (2025). The AI Continent Action Plan. https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan [15.04.2026].

European Commission (2026). Union of Skills. https://employment-social-affairs.ec.europa.eu/policies-and-activities/skills-and-qualifications/union-skills_en [15.04.2026].

European Parliament (2022). Decision (EU) 2022/2481 of the European Parliament and of the Council of 14 December 2022 establishing the Digital Decade Policy Programme 2030. In: Official Journal of the European Union L323., 4-26. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022D2481.

European Parliament (2023). European Declaration on Digital Rights and Principles for the Digital Decade. In: Official Journal of the European Union C23, 1-7. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32023C0123(01).
European Union (2026). Digital Skills and Jobs Platform - National Initiatives. https://digital-skills-jobs.europa.eu/en/initiatives/national-initiatives [15.04.2026].
Expertenkommission Forschung und Innovation (2023). Gutachten zu Forschung, Innovation und technologischer Leistungsfähigkeit Deutschlands 2023. Berlin. https://www.e-fi.de/fileadmin/Assets/Gutachten/2023/EFI_Gutachten_2023.pdf [30.06.2026].
Ferrari, A./Punie, Y./Redecker, C. (2012) Understanding Digital Competence in the 21st Century: An Analysis of Current Frameworks. In: Ravenscroft, A./Lindstaedt, S./Delgado Kloos, C./Hernández-Leo, D. (Eds.). 21st Century Learning for 21st Century Skills. Heidelberg. 79 - 92.
fit4internet (2026a). fit4internet: Über uns. Wien. https://www.fit4internet.at/view/verein [14.04.2026].
fit4internet (2026b). DIGITAL SKILLS BAROMETER 2025/2026. Wien. https://f4i.hbox.at/s/ReWWEAQKHTmWFYo?opendetails= [14.04.2026].
Mattar, J./Ramos, D.K./Lucas, M.R. (2022). DigComp-Based Digital competence Assessment Tools: Literature Review and Instrument Analysis. In: Education and Information Technologies 27 (8), 10843-10867. https://doi.org/10.1007/s10639-022-11034-3.
Newton, P.E. (2007). Clarifying the purposes of educational assessment. In: Assessment in Education: Principles, Policy & Practice 14(2), 149–170. https://doi.org/10.1080/09695940701478321
Oberländer, M./Beinicke, A./Bipp, T. (2020). Digital competencies: A review of the literature and applications in the workplace. In: Computers & Education 146, 103752. https://doi.org/10.1016/j.compedu.2019.103752.
Rathgeb, T./Doh, M./Tremmel, F./Gerigk, Y. (2025). SIM-Studie 2024. Stuttgart. https://mpfs.de/app/uploads/2025/05/SIM_2024_PDF_barrierearm.pdf [15.04.2026].
Sadler, D.R. (1989). Formative assessment and the design of instructional systems. In: Instructional Science 18(2), 119–144. https://doi.org/10.1007/BF00117714.
Sadler, D.R. (2010). Beyond feedback: developing student capability in complex appraisal. In: Assessment & Evaluation in Higher Education 35(5), 535–550. https://doi.org/10.1080/02602930903541015.
Stobart, G. (2008). Testing times: The uses and abuses of assessment. London. https://doi.org/10.4324/9780203930502
Stürz, R./Schlude, A./Harles, D./Mendel, U./Stumpf, C. (2025). Das bidt-Digitalbarometer 2025. Munich. https://doi.org/10.35067/xypq-kn75 [14.04.2026].
Vuorikari, R./Kluzer, S./Punie, Y. (2022a). DigComp 2.2: The Digital Competence Framework for Citizens. Luxembourg. https://publications.jrc.ec.europa.eu/repository/handle/JRC128415 [14.04.2026].
Vuorikari, R./Jerzak, N./Karpinski, Z./Pokropek, A./Tudek, J. (2022b). Measuring Digital Skills across the EU: Digital Skills Indicator 2.0. Luxembourg, https://doi.org/10.2760/897803 [30.06.2026].
Vuorikari, R./Pokropek, A./Muñoz, J. (2026). Enhancing digital skills assessment: introducing compact tools for measuring digital competence. In: Technology, Knowledge and Learning 31 (1), 519–546. https://doi.org/10.1007/s10758-025-09825-x.
Wiliam, D. (2006). Reliability, validity, and all that jazz. In: Education 3–13, 29(3), 17–21. https://doi.org/10.1080/03004270185200311.
Zimmermann, E./Tomczyk, S. (2025). Using Social Media to Promote Life Skills Among Adolescents: A Debate on Opportunities, Challenges, and Implications for Health and Education. In: Journal of prevention 46 (2), 201–211. https://doi.org/10.1007/s10935-025-00826-1.

SUB-SESSION

Overcoming Key Challenges of Data Donations: Balancing Individual Rights and Collective Benefit

20.11.2026, 14:00-15:30

Language: English

Chairs: Philipp Hartl | University of Regensburg; Jörg Haßler | University of Oslo, Norway; Leonie Manzke | FAU Erlangen-Nuremberg; Elisabeth Schmidbauer | LMU Munich; Daniel Schnurr | University of Regensburg; Verena Tiefenbeck | FAU Erlangen-Nuremberg

Introduction and Session Overview

A central challenge of digital transformation is how the personal data generated by individuals in their everyday interactions with digital services and platforms can be mobilized for the common good without compromising individual autonomy, privacy, and data sovereignty. Data donations represent a promising approach for society and researchers across disciplines to gain access to such data through the direct participation of individuals who voluntarily contribute their data for societal benefit. In practice, however, data donation initiatives continue to face persistent challenges, including low public awareness and inadequate data quality, which prevent the approach from reaching its full potential.

In light of persistent challenges to data donations, this sub-session pursues the overarching question: *Under what conditions can the voluntary disclosure of personal data be made viable, rigorous, and scalable for the common good, and what trade-offs must researchers, regulators, and platforms navigate to get there?*

By treating data donation as a prototypical approach for reconciling individual data rights with collective value creation, the sub-session aims to contribute an integrative, cross-disciplinary perspective to the digital transformation research discourse, connecting conceptual, methodological, legal, and behavioral viewpoints that are often addressed in isolation.

What Are Data Donations?

Over the past decade, the concept of data donations has received growing scholarly and policy interest. Various disciplines have adopted different terms – data altruism, data philanthropy, data sharing – each emphasizing different characteristics of the practice and reflecting different assumptions. To enhance definitional clarity and establish a common foundation for interdisciplinary discourse on data donations, the DD4SC project proposed an integrative definition according to which data donations involve (1) the disclosure of data by individuals, (2) of data related to themselves, (3) based on voluntary and informed consent, (4) for the purpose of promoting the common good (Schnurr et al., 2026).

Persistent Challenges in Applying Data Donations

The European **regulatory landscape** has, until recently, enabled the process of altruistic data disclosure through the frameworks of the General Data Protection Regulation, the Data Governance Act, and the Data Act (European Union, 2022; Kranz et al., 2023). However, the currently proposed Digital Omnibus Regulation might substantially restrict researchers' data access rights in the future, and thus undermine the practical viability of data donations (D3I: Europe's Data Donation Infrastructure, 2026).

Even within a supportive legal framework, significant challenges persist, such as those regarding **public awareness and motivation**. Familiarity with data donation as a concept remains low: a representative survey of Swiss online users found that only one in ten were familiar with the practice (Pfiffner & Friemel, 2023). This can have a significant impact on participation rates in data donation studies, which is particularly consequential when representative samples are needed.

Beyond awareness, sustaining donation behavior over time raises deeper motivational questions (Fast & Schnurr, 2026). Personal and prosocial motives may crowd out one another (Ghaffar & Widjaja, 2023), and data donations

can be susceptible to the **public-good dilemma**. Because donated data is non-rivalrous and others are often difficult to exclude from benefiting from it (Jones & Tonetti, 2020; Prainsack, 2019), individuals may free-ride on others' contributions rather than bear the personal costs of donating (Rockenbach et al., 2020).

Further, data donations introduce distinctive **privacy challenges**. Because the benefits of donating data extend to society rather than solely to the donor, established privacy calculus models may not straightforwardly apply to this prosocial context (Alashoor et al., 2025; Carlsson Hauff & Nilsson, 2023; Hartl & Manzke, 2026). Additionally, **broad consent mechanisms** raise critical questions about the boundaries of informed consent (Wenzel & Scherer, 2023) and individuals' willingness to donate data (Hartl & Schnurr, 2025).
Finally, ensuring the **technical consistency and quality of donated data** remains a challenge despite regulatory efforts to improve individual access rights. While digital trace data can offer higher accuracy than self-reports (Ohme et al., 2021), it is susceptible to selection bias and non-response errors (Sen et al., 2021), and substantial heterogeneity persists in the data download packages provided by digital platforms (Boeschoten et al., 2022; van Driel et al., 2022; Hase et al., 2024; Manzke & Hartl, 2024).

Session Goals and Discussion Questions
Because these challenges span technical, legal, and behavioral domains, no single discipline can address them alone. This session therefore brings together expertise from various fields to advance our understanding of how they might be overcome. The session opens with a keynote by Professor Boeschoten on “The Balancing Act of Data Donations,” drawing on her extensive experience leading data donation studies. This is followed by a reflection on empirical and theoretical work on the persistent challenges of data donations and potential solutions, including lessons learned from the bidt-funded DD4SC project, to provide a foundation for the discussions that follow. A 45-minute interdisciplinary panel discussion, including open-floor questions and comments, then brings these threads together around questions that cut across research disciplines, including:

- What research questions can *only* be answered through donated data? Where does untapped potential remain, and why?
- Do data donations have sufficient merit as a method in their own right, or are they ultimately a workaround until more suitable forms of data access become available?
- Could open access to anonymized donated datasets, or opt-out models, be ethically reconciled with informed consent – and can such approaches work at scale?
- Based on current research insights, what actionable measures can researchers take to improve data donation processes, from recruitment and consent design to data quality assurance?
- Data donations currently operate within an enabling EU-level legal framework. With the proposed Digital Omnibus Regulation expected to complicate researcher access to data, what does this mean for the future of data donations as a research tool?
- What concrete steps can policymakers, regulators, and platform providers take to create conditions that make data donations more effective and scalable?
- What conceptual, methodological, and ethical lessons from data donation research are generalizable to the wider study of digital transformation, and where do different research disciplines most productively converge or diverge in how they approach the underlying trade-offs?
- Beyond its use as a data-collection method, what does the data donation case teach us more broadly about reconciling individual data rights with collective value creation?
- Which of these insights transfer to other mechanisms of citizen-centric data governance (such as data trusts or personal data stores)?

Panel Discussion:

- **Asst. Prof. Dr. Laura Boeschoten**, Utrecht University, Lead of the project “D3I: Digital Data Donation Infrastructure”
- **Dr. Jakob Ohme**, Weizenbaum Institute Berlin, Research Group Lead “Digital News Dynamics”, Co-Principal Investigator in the TikTok data donation study “Dein Feed, Deine Wahl” and the #DSA40 Collaboratory
- **Julian Kohne**, GESIS Leibniz Institute for the Social Sciences, Researcher in Computational Social Science
- **Prof. Dr. Valerie Hase**, University of Klagenfurt, Research Group Lead “Digital Media and Methods Lab”, Principal Investigator in the DFG-funded project “Integrating Data Donation in Survey Infrastructure”
- **Prof. Dr. Daniel Schnurr** and **Prof. Dr. Verena Tiefenbeck** will moderate the panel discussion and add key insights from the DD4SC project

References

Alashoor T, Keil M, Jiang Z, Saffarizadeh K (2025) Privacy concerns and data donations: Do societal benefits matter? MIS Quarterly 49(2):429–464

Boeschoten L, Ausloos J, Möller JE, Araujo T, Oberski DL (2022) A framework for privacy preserving digital trace data collection through data donation. Computational Communication Research 4(2):388–423

Carlsson Hauff J, Nilsson J (2023) Individual costs and societal benefits: The privacy calculus of contact-tracing apps. Journal of Consumer Marketing 40(2):171–180

D3I: Europe's Data Donation Infrastructure (2026). Urgent Call to Reject the Omnibus Regulation with Regard to the GDPR's Right of Access. https://datadonation.eu/community/omnibus-letter. Last accessed: 14.04.2026.

European Union (2022) Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act). Official Journal of the European Union L 152:1–44

Fast V, Schnurr D (2026) Data donations for digital contact tracing: Short- and long-term effects of monetary incentives. Information Systems Research 37(1):627–642

Ghaffar AM, Widjaja T (2023) Towards a theory to explain prosocial data disclosure – An explorative investigation of the antecedents of infection disclosure. In: Proceedings of the 44th International Conference on Information Systems (ICIS 2023), Hyderabad, India

Hartl P, Manzke L (2026) Information sensitivity contamination in sequential data donation requests. In: Proceedings of the 34th European Conference on Information Systems (ECIS 2026), Milan, Italy

Hartl P, Schnurr D (2025) Encouraging data donations: An experimental analysis of contextual factors and personal antecedents. In: Proceedings of the 46th International Conference on Information Systems (ICIS 2025), Nashville, USA

Hase V, Ausloos J, Boeschoten L, Pfiffner N, Janssen H, Araujo T, Carrière T, de Vreese CH, Haßler J, Loecherbach F, Kmetty Z, Möller J, Ohme J, Schmidbauer E, Struminskaya B, Trilling D, Welbers K, Haim M (2024) Fulfilling data access obligations: How could (and should) platforms facilitate data donation studies? Internet Policy Review 13(3):1–37

Jones CI, Tonetti C (2020) Nonrivalry and the economics of data. American Economic Review 110(9):2819–2858

Kranz J, Kuebler-Wachendorff S, Syrmoudis E, Grossklags J, Mager S, Luzsa R, Mayr S (2023) Data portability. Business & Information Systems Engineering 65(5):597–607

Manzke L, Hartl P (2024) Data Donation for Impactful Insights: A Framework for Platform Selection and its Application to the Use Case of German Loyalty Card Providers. https://doi.org/10.13140/RG.2.2.15292.14723

Ohme J, Araujo T, de Vreese CH, Piotrowski JT (2021) Mobile data donations: assessing self-report accuracy and sample biases with the iOS screen time function. Mobile Media & Communication 9(2):293–313

Pfiffner N, Friemel TN (2023) Leveraging data donations for communication research: Exploring drivers behind the willingness to donate. Communication Methods and Measures 17(3):227–249
Prainsack B (2019) Data donation: How to resist the Leviathan. In: Krutzinna J, Floridi L (eds) The Ethics of Medical Data Donation. Springer, Cham, pp 9–22
Rockenbach B, Sadrieh A, Schielke A (2020) Providing personal information to the benefit of others. PLoS ONE 15(8):e0237183
Schnurr D, Hartl P, Manzke L, Schmidbauer E, Haßler J, Tiefenbeck V (2026) Data donations: data disclosure for the common good. Business & Information Systems Engineering. https://doi.org/10.1007/s12599-026-00989-z
Sen I, Flöck F, Weller K, Weiß B, Wagner C (2021) A total error framework for digital traces of human behavior on online platforms. Public Opinion Quarterly 85(S1):399–422
van Driel II, Giachanou A, Pouwels JL, Boeschoten L, Beyens I, Valkenburg PM (2022) Promises and pitfalls of social media data donations. Communication Methods and Measures 16(4):266–282
Wenzel C, Scherer A (2023) Sharing data for social good: From uninformed consent to misinformed dissent. In: Proceedings of the 50th Annual European Marketing Academy Conference (EMAC), Madrid, Spain

SUB-SESSION

Strafbare Deepfakes? Antworten auf neue Missbrauchsphänomene

20.11.2026, 14:00-15:30

Sprache: Deutsch

Chairs: Lorenz Meinen | Universität Bayreuth; Christian Rückert | Universität Bayreuth; Astrid Schomäcker | Universität Bayreuth; Lena Kästner | Universität Bayreuth; Niklas Kühl | Universität Bayreuth

Thema der Sub-Session

Deepfakes sind ein relativ junges Phänomen. Trotz des erheblichen Missbrauchspotenzials sind sie daher bislang nur wenig reguliert. Wegen der potenziell großen Schäden für Betroffene wird vor allem die strafrechtliche Sanktionierung kontrovers diskutiert. Hier fokussieren sich die Fachkreise vermehrt auf sexualisierende Darstellungen, denen das deutsche Strafrecht bislang kaum etwas entgegenzusetzen hat. Ob es hier neue strafrechtliche Sanktionen braucht und welches Unrecht diese erfassen sollten, sind komplexe Fragen der Strafrechtstheorie, die nur unter Rückgriff auf rechtswissenschaftliche und philosophische Überlegungen beantwortet werden können.
Bisher ist noch nicht abschließend geklärt, welche Verhaltensweisen in Bezug auf sexualisierende Deepfakes strafbar sein sollten, was auch daran liegt, dass bislang noch unklar ist, welche Verhaltensweisen schon heute strafbar sind. Um hier tragfähige Antworten zu finden, werden nicht nur Einblicke aus den Rechtswissenschaften und der Philosophie benötigt, da über gesetzliche Vorhaben nur im demokratischen Prozess entschieden werden kann. Es ist daher notwendig, in den Austausch mit den demokratischen Repräsentanten zu treten, um so Lösungen zu finden, die nicht nur juristisch-philosophisch sinnvoll sind, sondern gleichzeitig auch demokratisch umsetzbar und mehrheitsfähig sind.

Neben den Fragen der Rechtssetzung, dürfen aber auch Fragen der Strafverfolgung nicht vernachlässigt werden, da neue Strafnormen nur im Umfang ihrer Durchsetzbarkeit die bezweckten Effekte erzielen können. Ein holistischer Ansatz kann daher nur unter Einbeziehung der Möglichkeiten und Grenzen der Strafverfolgung im digitalen Raum gelingen, da hier der Großteil aller Deepfakes verbreitet wird. Aus diesem Grund finden sich in der Sub-Session „Strafbare Deepfakes? Antworten auf neue Missbrauchsphänomene“ Vertreter*innen der Rechtswissenschaften, der Philosophie, der Politik und der Strafverfolgung zusammen, um herauszuarbeiten, welche Rechtsgüter sexualisierende Deepfakes betreffen, ab wann die Grenze zur Strafwürdigkeit überschritten ist und welche Effekte von der strafrechtlichen Sanktionierung zu erwarten sind.
Den Auftakt der Sub-Session bilden kurze Impulsvorträge der Referent*innen, auf deren Grundlage eine von Professor Dr. Rückert moderierte Diskussion stattfindet, in der die verschiedenen Perspektiven auf das Phänomen der sexualisierenden Deepfakes zusammengeführt und kritisch beleuchtet werden sollen.

Impulsvorträge und Panel-Diskussion:

Christoph Ebert | Bayerisches Staatsministerium der Justiz

Brian Valerius | Universität Passau

Um einen strafrechtlichen Reformbedarf im Zusammenhang mit Deep Fakes beurteilen zu können, ist zunächst die Strafbarkeit nach geltendem Recht zu untersuchen. Insoweit bleibt festzuhalten, dass ein eigenständiger Straftatbestand, der allein und gezielt das Phänomen „Deep Fakes“ in den Blick nimmt, bislang nicht existiert. Dies ist allerdings bei der grundsätzlich an Rechtsgütern orientierten Strafgesetzgebung auch nicht unbedingt

verwunderlich. Problematisch erscheint hingegen, dass es zwar durchaus ein beachtliches Spektrum von Straftatbeständen gibt, die den missbräuchlichen Umgang mit Deep Fakes sanktionieren können. Die meisten Delikte stehen indessen nur im Raum, wenn Deep Fakes zu bestimmten (z.B. betrügerischen oder erpresserischen) Zwecken verwendet werden oder wenn sie bestimmte, per se rechtswidrige (z.B. ehrverletzende, volksverhetzende oder kinder- und jugendpornographische) Inhalte aufweisen.
Die Verletzung von Persönlichkeitsrechten, die häufig mit der Herstellung und Verbreitung von Deep Fakes einhergeht, ist hingegen nur selten strafbar. Vor allem die hierfür prädestinierten §§ 201, 201a StGB, die letztlich das Recht am eigenen Wort und das Recht am eigenen Bild schützen, vermögen aus verschiedenen Gründen in der Regel weder die Herstellung noch die Verbreitung von Deep Fakes zu erfassen. Somit kommt oftmals allenfalls eine Strafbarkeit nach § 33 des aus dem Jahr 1907 stammenden sog. Kunsturheberrechtsgesetzes in Betracht. Dieser Paragraph stellt indessen nur das Verbreiten (oder öffentliche Zurschaustellen) eines Bildnisses, d.h. einer Abbildung, auf der sich eine bestimmte Person erkennen lässt, unter Strafe und droht hierfür bis zu ein Jahr Freiheitsstrafe oder Geldstrafe an.

Astrid Schomäcker | Universität Bayreuth

Die Frage, ob bestimmte Handlungen -in unserem Fall die Erstellung und Verbreitung von Deepfakes- durch den Gesetzgeber bestraft werden sollten, ist im Kern eine ethische und rechtsphilosophische. Um sie zu beantworten, muss man zwischen Handlungen unterscheiden, die moralisch problematisch sind, und solchen, die (straf-)rechtlich verfolgt werden sollten. Dieser Impuls arbeitet zu diesem Zweck Fallbeispiele heraus und diskutiert diese mit Blick sowohl auf ihre ethischen Dimensionen als auch auf die Strafwürdigkeit. Diskutiert werden dabei die Fragen, welche Schäden durch die Erstellung und Verbreitung von pornographischen Deepfakes entstehen, ob dabei bestimmte Rechte verletzt werden (z.B. Privatsphäre, Autonomie, informationelle Selbstbestimmung) und an welchem Punkt eine strafrechtliche Verfolgung moralisch geboten erscheint.

Maria Gahn | Universität des Saarlandes

Der Gesetzentwurf des Bundesministeriums für Justiz und Verbraucherschutz zielt darauf ab, strafrechtliche Lücken in Bezug auf Deepfakes zu schließen und zugleich die EU-Gewaltschutzrichtlinie umzusetzen. Von dem Entwurf werden nicht-sexualisierte rufschädigende Deepfakes (§ 201b StGB-E) sowie sexualisierte Deepfakes von Erwachsenen (§ 184k Abs. 1 Nr. 4 StGB-E) und Minderjährigen (§ 184b Abs. 1 S. 1 Nr. 3 StGB-E, § 184c Abs. 1 S. 1 Nr. 3 StGB-E) erfasst. Nicht erfasst werden KI-generierte Inhalte ohne Personenbezug. Besonders umstritten ist die Regelung zu sexualisierten Deepfakes Erwachsener. Diese steht im Zentrum dieses Beitrags. Untersucht werden die Anforderungen an das Tatobjekt der Bildaufnahmen, welche auf sexualisierte Handlungen oder bestimmte Körperstellen begrenzt sind, und die Frage, wann eine Manipulation den erforderlichen Anschein einer tatsächlichen Aufnahme erweckt. Zu diskutieren ist weiter die Ausgestaltung der Tathandlungen. Nach dem Entwurf wird insbesondere bereits das Herstellen erfasst, was der Gefahr des unberechtigten Zugriffs Dritter sowie Beweisproblematiken begegnet; teils wird diese Ausgestaltung indes als zu weite Vorverlagerung der Strafbarkeit kritisiert. Ziel dieses Beitrags ist es, die Reichweite und die wesentlichen Wertungsfragen der geplanten Regelung herauszuarbeiten.

Lorenz Meinen | Universität Bayreuth

Im Rahmen des durch das bidt geförderten Projektes FoGG (For the Greater Good - Deepfakes in der Strafverfolgung), ist ein Entwurf für die Strafbarkeit von Deepfakes entstanden. Dieser basiert auf einer Analyse der bisher bekannt gewordenen Entwürfe und der daraus folgenden Erkenntnis, dass für jeden dieser Entwürfe viel spricht, sie aber dennoch voneinander lernen können. Das Ergebnis ist ein § 201b StGB-E, der die Zugänglichmachung bestimmter (Deepfake-) Darstellungen an Dritte unter Strafe stellt.
Im Strafrecht ist es besonders wichtig, gut verständliche Tatbestandsmerkmale zu schaffen, damit jede*r verstehen kann, welches Verhalten strafbar ist. Gleichzeitig dürfen diese Merkmale nicht allzu ausufernd gestaltet sein, um sicherzustellen, dass nur wirklich

strafwürdige Verhaltensweisen sanktioniert werden. Unter Beachtung dieser Grundsätze sollen im Entwurf des FoGG-Projektes sowohl sexualisierende als auch sonstige Darstellungen sanktioniert werden. Die Strafbarkeit von sexualisierenden Deepfakes wird an die Abbildung bestimmter Motive geknüpft, damit die strafbaren Handlungen klar bestimmt sind. Jenseits der sexualisierenden Darstellungen können Deepfakes auch erhebliche Schäden verursachen, wenn sie dazu verwendet werden, Personen bestimmte Verhaltensweisen oder Äußerungen zuzuschreiben. Der Entwurf setzt daher wirklichkeitsnahe Zuschreibungen mittels Deepfakes unter Strafe, wenn diese mit Schädigungsabsicht weitergegeben werden. Der Beitrag stellt diesen Entwurf und seine Hintergründe kurz vor, damit er im Spiegel der anderen thematisierten Entwürfe diskutiert werden kann.

Markus Hartmann | Leitender Oberstaatsanwalt

Mit der Digitalisierung aller Bereiche des wirtschaftlichen, sozialen und politischen Lebens geht auch eine Digitalisierung der Kriminalität einher. Der Staat hat auf diese Entwicklung reagiert, etwa durch die Einrichtung von Cybercrime-Zentralstellen in den Bundesländern, d. h. spezialisierten Einheiten der Justiz, die die staatsanwaltschaftlichen Aufgaben der Strafverfolgung mit besonderer thematischer und vor allem technischer Expertise wahrnehmen. Dieser Fokus auf herausgehobene Erscheinungsformen digitaler Kriminalität ist wichtig und richtig. Gleichzeit darf nicht übersehen werden, dass digitale Tatbegehungsweisen mittlerweile alle Kriminalitätsfelder durchdringen: Offline-Betrug ist vielfach von Online-Abzocke verdrängt, die Beleidigung am Gartenzaun findet sich mittlerweile vornehmlich in den sozialen Medien wieder.
Nicht zuletzt der Bereich der Straftaten des 13. Abschnitts des Besonderen Teils des Strafgesetzbuchs ist von der umfassenden Verlagerung ins Digitale geprägt. Kindesmissbrauchsdarstellungen werden online verbreitet. Der Kontext der übergroßen Zahl von tatsächlichen Missbrauchshandlungen ist internetkonnex. Die Möglichkeiten der Bildmanipulation und der Generierung komplett künstlicher Inhalte durch Künstliche Intelligenz – insbesondere sogenannte Deepfakes – erweisen sich als Treiber dieser Entwicklung und wirken dabei über den Kernbereich der Sexualstraftaten hinaus in vielfältige Rechts- und Praxisfragen hinein. Deepfakes stellen die Strafverfolgung damit vor besondere und neue Herausforderungen. Debattiert werden aktuell vor allem Veränderungen des materiellen Strafrechts. Sie sind zwingend zu flankieren durch den Blick auf die tatsächlichen Möglichkeiten der Rechtsdurchsetzung im Internet. Der Beitrag beleuchtet, was neben neuen Strafgesetzen nötig ist, um sexualisierte Deepfakes zu bekämpfen.

SUB-SESSION

Führung und Organisationen im digitalen Wandel: Chancen, Herausforderungen und KI-Perspektiven

20.11.2026, 15:45-17:15

Sprache: Deutsch

Chairs: Melanie Vilser | Universität der Bundeswehr München; Maxi Holzapfel | Munich Business School; Ellen Schmid | Universität der Bundeswehr München; Markus Böhm | Hochschule Landshut

Die fortschreitende Digitalisierung und der Einzug Künstlicher Intelligenz (KI) verändern das Fundament moderner Arbeitswelten. Diese Sub-Session widmet sich der kritischen, interdisziplinären Reflexion dieser Dynamiken und ihren konkreten Auswirkungen auf die Führung. Aufbauend auf den Ergebnissen des World Cafés „Führung in digitalen Arbeitswelten“ des bidt-Drittmittelprojekts ELDiCo (Exploitative Leadership in Digital Collaboration) setzt die Session neue Impulse für Forschung und Praxis.
Ziel ist es, zentrale Fragestellungen der digitalen Transformationsforschung zu diskutieren, neue Forschungsfragen zu generieren und praxisorientierte Handlungsperspektiven abzuleiten. Dabei wird das Phänomen aus einer systemischen 360-Grad-Perspektive betrachtet und die Teilnehmenden systematisch von der mikroorganisationalen Ebene der individuellen Belastung hin zur makroorganisationalen Systemebene geführt.

Impulsvorträge & Roundtable-Diskussionen:

(Negatives) Führungsverhalten in digitalen Arbeitswelten

Ellen Schmid | Universität der Bundeswehr München

Dieser Panel-Talk widmet sich den destruktiven Schattenseiten und neuen Manifestationen negativer Führung im virtuellen Raum, um das notwendige Problembewusstsein für technologische Entlastungspotenziale zu schaffen. Im Zentrum der wissenschaftlichen und praxisorientierten Diskussion steht die Frage, inwiefern die fortschreitende Digitalisierung die konkrete Manifestation destruktiven Führungsverhaltens verändert. Dabei werden neuartige, technologieinduzierte Verhaltensmuster – wie digitales Mikromanagement oder der Druck zur ständigen Erreichbarkeit – analysiert und traditionellen Face-to-Face-Verhaltensweisen gegenübergestellt. Ein besonderer Fokus liegt auf der ambivalenten Rolle der räumlichen Distanz in virtuellen Teams: Es wird kritisch untersucht, ob diese geografische Trennung eher als Puffer und somit als Schutzraum für Mitarbeitende wirkt oder ob sie durch schwindende soziale Hemmschwellen, insbesondere in der schriftlichen Textkommunikation, als Katalysator für destruktives Verhalten fungiert.

KI als Entscheidungshilfe für Führungskräfte?

Markus Böhm | Hochschule Landshut

Aufbauend auf den zuvor diskutierten Belastungen setzt dieser Panel Talk den konstruktiven Gegenpol, indem er das Potenzial Künstlicher Intelligenz (KI) zur technologischen Entlastung im Management untersucht. Im Zentrum steht die Analyse, wie KI als intelligenter Copilot Führungskräfte bei komplexen Entscheidungsfindungen unterstützen und durch gezielte administrative Entlastung den negativen Dynamiken virtueller Führung aktiv entgegenwirken kann. Gleichzeitig wirft dieser technologiegestützte Wandel fundamentale Fragen zur Veränderung des gesamten organisationalen Gefüges auf. Kritisch beleuchtet wird dabei insbesondere die feine Demarkationslinie zwischen gesunder, datengestützter Entscheidungsunterstützung und dem schleichenden Verlust genuin menschlicher Führungskompetenzen, ausgelöst durch eine unkritische Übertragung von Verantwortung auf Algorithmen (Automation Bias).

Relationaler Wandel und soziale Dynamiken im KI-Zeitalter

Julia Reif | Universität der Bundeswehr München

Als logische Konsequenz des veränderten Zusammenspiels von Mensch, Technologie und digitaler Führung führt dieser finale Panel Talk die vorherigen Perspektiven auf einer Systemebene zusammen. Im Zentrum steht die fundamentale Frage, wie dezentrale KI-Systeme als neue, autonome Interaktionspartner die relationalen Modelle, Kommunikationsmuster und sozialen Beziehungen innerhalb von Organisationen beeinflussen. Es wird kritisch untersucht, wie sich traditionelle Beziehungsgefüge und etablierte Interaktionsstrukturen verändern, wenn der exklusive Zugang zu entscheidungsrelevanten Informationen nicht mehr an personelle Wissensträger gebunden ist, sondern durch Technologie demokratisiert wird. Die Diskussion analysiert die resultierenden Dynamiken in der Unternehmenskultur, die sich im spannungsreichen Feld zwischen digitaler Autonomie, Beziehungsqualität und technologischer Kontrolle bewegen, um daraus die notwendigen strukturellen Leitplanken für die Praxis abzuleiten.

SUB-SESSION

KI statt Streife und Richterbank? Gesellschaftliche Akzeptanz und Vertrauen in den Einsatz digitaler Technologien bei Polizei und Justiz

20.11.2026, 15:45-17:15

Sprache: Deutsch

Chairs: Andreas Jungherr | bidt & Otto-Friedrich Universität Bamberg; Bruno Albert | bidt

Digitalisierung und digitale Technologien verändern die Gesellschaft nachhaltig und haben Auswirkungen in zahlreichen Bereichen nutzbar. Dies gilt auch für die Bereiche Polizei und Justiz. Sie versprechen zum einen Potenziale hinsichtlich der Steigerung von Effizienz, Transparenz und Zugänglichkeit (Martini, 2017), gleichzeitig werden aber auch grundlegende Fragen nach den gesellschaftlichen Auswirkungen und Rahmenbedingungen der behördlichen Digitalisierung aufgeworfen (Will-Zocholl, 2023).
Die Anwendung digitaler Technologien im Bereich Polizei und Justiz ist dabei besonders brisant, da der Einsatz solcher Technologien weitreichende Konsequenzen für Betroffene haben könnte. Das gilt insbesondere dann, wenn nicht nur bestehende Prozesse digitalisiert werden, etwa durch das Ersetzen von Papierakten durch digitale Akten, sondern wenn neue Technologien auch neue Abwägungen erforderlich machen, beispielsweise zwischen Datenschutz und Effizienz.

Bei diesen gesellschaftlichen Fragen ist es von großer Wichtigkeit, die Einstellungen in der Bevölkerung zum Einsatz digitaler Technologien in diesen Bereichen zu erfassen und zu verstehen. Aufgrund der sehr schnellen technischen Entwicklung in Bereichen wie der künstlichen Intelligenz verschieben sich auch die Grenzen des technisch Möglichen mit großer Geschwindigkeit. Daraus ergibt sich ein Bedarf an aktuellen empirischen Analysen, welche die gesellschaftlichen Einstellungen zu einem möglichen Einsatz dieser Technologien untersuchen.
Die Subsession befasst sich daher mit dem gesellschaftlichen Blick auf den möglichen Einsatz digitaler Technologien in Polizei und Justiz. Ein besonderes Augenmerk wird dabei auf den Einfluss von Institutionenvertrauen auf die Einstellungsbildung gelegt. Hierbei werden Perspektiven aus Politikwissenschaft, Soziologie, Psychologie und der Rechtswissenschaft integriert.

Der erste Vortrag präsentiert Ergebnisse einer länderübergreifenden Befragung zur gesellschaftlichen Einschätzung des Einsatzes künstlicher Intelligenz durch die Polizei. Der zweite Vortrag stellt Ergebnisse eines faktoriellen Survey-Experiment zur technischen und rechtlichen Ausgestaltung einer konkreten polizeilichen Kommunikationsüberwachungsmaßnahme vor. Der dritte Vortrag präsentiert Ergebnisse aus drei Vignettenbefragungen zum Einfluss generativer KI-Nutzung in der richterlichen Urteilsfindung sowie der emotionalen Fallkomplexität auf die wahrgenommene Vertrauenswürdigkeit von Richtern und die Akzeptanz richterlicher Urteile.
Die Subsession leistet damit einen Beitrag zu einem tieferen Verständnis der gesellschaftlichen Bedingungen, unter denen digitale Technologien in sensiblen staatlichen Handlungsfeldern gesellschaftlich akzeptiert werden.

Impulsvorträge:

Andreas Jungherr | bidt; Roland A. Stürz | bidt; Antonia Schlude | bidt

Künstliche Intelligenz verändert die Polizeiarbeit rasant. Doch diese Instrumente werfen ein grundlegendes demokratisches Dilemma auf: Dieselben Bürger:innen, die von mehr Sicherheit profitieren könnten, sind gleichzeitig einer verstärkten algorithmischen Überwachung ausgesetzt. In diesem Beitrag wird untersucht, wie die Menschen in neun Ländern – China, Finnland, Frankreich, Deutschland, Italien, Spanien, Taiwan, dem Vereinigten Königreich und den Vereinigten Staaten – mit diesem Zielkonflikt umgehen.

Wir stellen die These auf, dass zwei kontextuelle Faktoren diese Bewertung strukturieren. Die Wahrnehmung von Kriminalität verstärkt die Bedeutung der Sicherheitsvorteile, während institutionelles Vertrauen die wahrgenommenen Kosten der algorithmischen Unterwerfung verringert. Entscheidend ist, dass diese Faktoren miteinander interagieren sollten: Die Sorge um Kriminalität sollte stärker mit der Unterstützung für KI in der Polizeiarbeit verbunden sein, wenn das Vertrauen in rechtliche Institutionen hoch ist, und Vertrauen sollte am wichtigsten sein, wenn die wahrgenommenen Sicherheitsbedrohungen akut sind. Wir erwarten ferner, dass allgemeiner KI-Optimismus und -Skeptizismus die Grundunterstützung prägen, während KI-Wissen die Abhängigkeit von diesen allgemeinen Vorurteilen verringert.

Wir überprüfen diese Erwartungen anhand länderübergreifender Umfragedaten aus dem bidt digitalbarometer.international 2026 mit rund 1.000 Befragten pro Land. Die Studie leistet einen Beitrag zur Forschung über algorithmische Governance, indem sie zeigt, dass die öffentliche Unterstützung für KI in staatlichen Zwangskontext nicht als einfache Technologieakzeptanz verstanden werden kann. Stattdessen spiegelt sie wider, wie Bürger:innen das Versprechen kollektiver Sicherheit gegen Risiken für Freiheit, Fairness und demokratische Rechenschaftspflicht abwägen.

Bruno Albert | bidt; Joanna Klauser | bidt

Dieser Beitrag untersucht den Einfluss der rechtlichen und technischen Ausgestaltungsmöglichkeiten eines polizeilichen Eingriffs zur Überwachung der Telekommunikation auf die gesellschaftliche Akzeptanz mithilfe eines faktoriellen Survey-Experiments mit 1.200 Befragten in Deutschland. Das Experiment wurde im Rahmen eines interdisziplinären Forschungsprojekts mit Beteiligten aus den Rechtswissenschaften, der Informatik und den Sozialwissenschaften entwickelt.

Das faktorielle Experiment verwendet ein 3x2x2x2x2x3 full-factorial Design (Auspurg und Hinz, 2015). Jeder Befragte bewertete drei Vignetten, was zu über 3.600 bewerteten Vignetten führt. Dabei werden folgende Attribute einer fiktiven Überwachungsmaßnahme manipuliert, um ihren Einfluss auf die gesellschaftliche Akzeptanz zu messen: (1) der Ermittlungsanlass, (2) die Möglichkeit zur Massenüberwachung, (3) die Art der überwachten Daten, (4) institutionelle Kontrolle durch Gerichte, (5) der technische Zugang zum Gerät und (6) der Einsatz künstlicher Intelligenz im Überwachungsprozess.
Der Beitrag untersucht darüber hinaus den Einfluss von Institutionenvertrauen auf die Akzeptanz der Überwachungsmaßnahmen und geht dabei der Frage nach, ob ein „blindes Vertrauen“ gegenüber der Polizei existiert, das zu einer kritiklosen Akzeptanz staatlicher Überwachungsmaßnahmen führt, unabhängig von deren Ausgestaltung und institutioneller Kontrolle.

Marvin Walczok | LMU München; Friederike Funk | LMU München

Generative künstliche Intelligenz (genAI) könnte die richterliche Entscheidungsfindung verändern, indem sie Richter:innen unterstützt oder ersetzt. Frühere Untersuchungen zeigen, dass Laien menschlichen Richtern mehr Vertrauen entgegenbringen und sie eher akzeptieren als algorithmische Richter. Es ist jedoch wenig über die öffentliche Wahrnehmung hybrider Entscheidungsfindung – also menschlicher Richter:innen mit Unterstützung durch genAI – bekannt, die das realistischste Szenario für die Justiz weltweit in der Zukunft darstellen wird. Aufbauend auf dem Modell des organisatorischen Vertrauens, der maschinellen Heuristik und dem Capability-Personalization Framework untersuchen wir den Einfluss der Art des Entscheidungsträgers auf die wahrgenommene Vertrauenswürdigkeit (Kompetenz, Wohlwollen und Integrität) des Entscheidungsträgers sowie auf die Akzeptanz von Urteilen. Wir stellen die Hypothese auf, dass bei menschlicher und hybrider richterlicher Entscheidungsfindung eine höhere Kompetenz, ein höheres Wohlwollen und eine höhere Akzeptanz sowie eine geringere Integrität zu verzeichnen sind als bei genAI. In drei vorab registrierten Online-Experimenten mit deutschen Erwachsenen aus Convenience-Sampling und Panel-Anbietern (PsyWeb, Prolific) präsentierten wir hypothetische Straftaten mit anschließenden Gerichtsverfahren und Urteilen. Wir manipulierten die Art des Entscheidungsträgers als Zwischenfaktor mit vier Stufen: (1) menschlicher Richter, (2)

Hybrid mit hohem menschlichem Anteil, (3) Hybrid mit geringem menschlichem Anteil, (4) GenAI. Die Ergebnisse von Experiment 1 (N1 = 323) zeigen eine Reduktion bei Kompetenz, Wohlwollen, Integrität und Akzeptanz mit abnehmender menschlicher Beteiligung. Die Experimente 2 (N2 = 674) und 3 (N3 = 772) ergeben ähnliche Bewertungen hinsichtlich Vertrauenswürdigkeit und Akzeptanz für menschliche und hybride Bedingungen, bei denen die Bewertungen signifikant höher waren als bei GenAI. Die Ergebnisse liefern gemischte Belege für die Wahrnehmung hybrider richterlicher Entscheidungsfindung, stützen jedoch die Präferenz für menschliche Richter. Insgesamt warnen sie vor Einbußen bei Vertrauenswürdigkeit und Akzeptanz, falls GenAI Richter:innen ersetzen statt unterstützen würde, heben jedoch das Potenzial hybrider richterlicher Entscheidungsfindung hervor.

bidt Digital Transformation Research Conference

Referent:innen-Liste / List of Presenters

Name	Affiliation	Seite / Page
Absmayr, Thomas	FHV – Vorarlberg University of Applied Sciences, Austria	106
Albert, Bruno	bidt	137
Appel, Markus	University of Würzburg	89
Arda, Hakan	SHIFT	26
Ayeh, Diana	Helmholtz Centre for Environmental Research	97
Balta, Dian	fortiss	78
Baumüller, Imme	Unabhängige Forscherin und Unternehmensberaterin, Digitale Transformation	58
Biedermann Camposano, Sabine	Harz University of Applied Sciences & TU Munich	97
Blank, Marco	FAU Erlangen-Nürnberg	116
Bleicher, Alena	Harz University of Applied Sciences	97
Boeschoten, Laura	Utrecht University	129
Böhm, Markus	Hochschule Landshut	134
Bredée, Jülide	Capgemini Invent	76
Brenner, Barbara	Universität St. Gallen, Schweiz	57
Brenner, Walter	Universität St. Gallen, Schweiz	57
Brodbeck, Felix C.	LMU Munich	32
Brunner, Christina M.	LMU Munich	32
Butollo, Florian	Goethe-Universität Frankfurt	115
Castillo Rivadeneira, Juan Camilo	InfAI	49
Cosgrove, Jude	Joint Research Centre, European Commission	123
Crompton, Laura	TH Ingolstadt	37
Czesnick, Hjördis	OWID	114
Doh, Michael	Catholic University of Applied Sciences Freiburg	123
Domany-Funtan, Ulrike	Secretary General fit4internet, Austria	124
Domokos, Daniella	Capgemini Invent	76
Doseva, Steliyana	bidt	118
Doubali, Benjamin	Johannes Gutenberg Universität Mainz	53
Ebert, Christoph	Bayerisches Staatsministerium der Justiz	131
Elkin-Koren, Niva	Harvard University, USA	86
Fourie, Willem	School for Data Science and Computational Thinking, Stellenbosch University, South Africa	121
Frieß, Dennis	Heine Center for Resilient Democracy (HeiCReD), Heinrich-Heine-Universität Düsseldorf	119
Funk, Friederike	LMU München	137

Gahn, Maria	Universität des Saarlandes	132
Garfinkel, Simson	BasisTech, USA	8
Glaser, Julius	Georg-August-Universität Göttingen	18
Göbl, Linda	Catholic University of Applied Sciences Freiburg	123
Graf, Angela	bidt	116
Gründler, Anna	deepset	87
Guio Espanol, Armando	Berkman Klein Center for Internet & Society, Harvard University, USA	82
Hajek, Katharina V.	LMU München	118
Händel, Marion	Ansbach University of Applied Sciences	75, 114
Hardwig, Thomas	Georg-August-Universität Göttingen	18
Harles, Danilo	bidt	124
Hartmann, Markus	Leitender Oberstaatsanwalt	133
Hase, Valerie	University of Klagenfurt	129
Heckmann, Dirk	bidt & TU München	114
Herz, Marie	SHIFT	26
Jarrahi, Ensieh	Independent Scholar	72
Jarrahi, Mohammad Hossein	University of North Carolina at Chapel Hill, USA	72
Jungherr, Andreas	bidt / Universität Bamberg	118, 136
Kalpokas, Ignas	Vytautas Magnus University, Lithuania	64
Karami, Amir	Kennesaw State University, USA	72
Kästner, Lena	Universität Bayreuth	114
Katzinski, Ann	WZB & Weizenbaum-Institut	115
Kiemer, Anna	bidt	14
Klauser, Joanna	bidt	137
Kohne, Julian	GESIS	129
Kolhep, Paula	University of Passau	29
Kravets-Meinke, Daria	bidt	89
Krönke Lieb, Christoph	Universität Bayreuth	87
Krzywdzinski, Martin	WZB/Weizenbaum-Institut/Helmut-Schmidt-Universität Hamburg	116
Kugler, Katharina G.	LMU Munich	32
Kühl, Niklas	University of Bayreuth	8
Kulla, Marlene	WZB/Weizenbaum-Institut	115
Lenk, Hannah	University of Bamberg	78
Lutz, Christoph	BI Norwegian Business School, Norway	72, 99
Meinen, Lorenz	Universität Bayreuth	132
Messingschlager, Tanja	University of Würzburg	89
Müller, Nicholas H.	SHIFT	26
Müßig, Michael	SHIFT	26
Naujoks-Schober, Nick	Ansbach University of Applied Sciences	75

Newlands, Gemma	University of Zurich, Switzerland	99
Nürnberger, Ruth	Brückner Maschinenbau	21
Oberländer, Anna Maria	Universität Bayreuth	87
Oermann, Markus	SHIFT	26
Ohliger, Ursula Alexandra	Deutsches Forschungszentrum für Künstliche Intelligenz	119
Ohme, Jakob	Weizenbaum Institute Berlin	129
Ostendorf, Marietta	Universität Potsdam	42
Ostern, Nadine	Queensland University of Technology, Australia	93
Ozornina, Nadezhda	Department of Media and Communication, LMU Munich	121
Park, Lisa	University of Alabama at Birmingham, USA	72
Partisch, Ina	Fraunhofer FIT / Frankfurt University of Applied Sciences	93
Pfeiffer, Sabine	bidt & FAU Erlangen-Nuremberg	8, 116
Pöhlmann, Kendra	OTH Regensburg	21
Prem, Erich	eutema, TU Vienna & European Digital Humanism Initiative	91
Pretschner, Alexander	bidt & TU Munich	91
Redepenning, Marc	University of Bamberg	78
Reif, Julia	Universität der Bundeswehr München	135
Rieger, Diana	Department of Media and Communication, LMU Munich	121
Roberts, Molly	University of California San Diego, USA	68
Römer, Vanita	InfAI	49
Rosemann, Michael	Queensland University of Technology, Australia	87, 93
Roth, Angela	FAU Erlangen-Nuremberg	78
Rüger, Korbinian	LMU Munich	111
Ruppenthal, Josefine	TH Ingolstadt	37
Safin, Damir	fortiss	78
Salzberger, Franziska L.	LMU Munich	32
Schlögl, Stephan	University of Passau	29
Schmid, Ellen	Universität der Bundeswehr München	134
Schmid, Ursula K.	LMU Munich	120
Schmid-Petri, Hannah	University of Passau & bidt	29, 89
Schnurr, Daniel	University of Regensburg	129
Schomäcker, Astrid	Universität Bayreuth	132
Schlude, Antonia	bidt	136
Schusterbauer, Johannes	Department of Computer Science, LMU Munich	121
Seferi, Fabio	IMT School for Advanced Studies Lucca & University of Florence, Italy	82
Sehne, Julian	SHIFT	26
Sengewald, Timon	FAU Erlangen-Nuremberg	78
Spielbauer, Philipp	LMU Munich	120
Staudacher, Klaus	bidt & European Digital Humanism Initiative	90
Streitbörger, Chiara	Universität Potsdam	42

Stürz, Roland A.	bidt	124, 136
Sundar, S. Shyam	The Pennsylvania State University, USA	88
Talgatova, Malika	TU Munich	68
Theocharis, Yannis	TU Munich	68
Tiefenbeck, Verena	FAU Erlangen-Nuremberg	129
Umbrello, Steven	Institute for Ethics and Emerging Technologies, University of Turin & European Digital Humanism Initiative	91
Valerius, Brian	Universität Passau	131
van Giffen, Benjamin	Universität Liechtenstein, Liechtenstein	57
Völzow, Christine	vbw – Vereinigung der Bayerischen Wirtschaft e.V.	87
Walczok, Marvin	LMU München	137
Wasserman Rozen, Hofit	Tel-Aviv University, Israel	86
Weidmann, Nils	University of Konstanz	68
Werner, René	Johannes-Kepler-Universität Linz, Österreich	11
Werther, Charlotte	Open Search Foundation e.V.	103
Yu, Shubin	HEC Paris, France	99
Zhou, Shijia	Center for Information and Language Processing (CIS), LMU Munich	120
Zierer, Antonia	Bayerisches Staatsministerium für Digitales	87

bidt Digital Transformation Research Conference

Conference Organisation

Conference Chairs

Felix Freiling | bidt & FAU Erlangen-Nürnberg
Sabine Pfeiffer | bidt & FAU Erlangen-Nürnberg

Organising Team

Felix Freiling | bidt & FAU Erlangen-Nürnberg
Sabine Pfeiffer | bidt & FAU Erlangen-Nürnberg
Angela Graf | bidt
Andreas Wenninger | bidt

Programme Committees

Wirtschaft & Arbeit / Economy & Work

Sabine Pfeiffer | bidt & FAU Erlangen-Nürnberg
Thomas Hess | bidt & LMU München
Alexander Pretschner | bidt & TU München
Angela Graf | bidt

Kommunikation & Gesellschaft / Communication & Society

Hannah Schmid-Petri | bidt & Universität Passau
Ute Schmid | bidt & Universität Bamberg
Andreas Jungherr | bidt & Universität Bamberg
Andreas Wenninger | bidt

Recht & Ethik / Law & Ethics

Julian Nida-Rümelin | bidt & LMU München
Dirk Heckmann | bidt & TU München
Felix Freiling | bidt & FAU Erlangen-Nürnberg
Niina Zuber | bidt

Sub-Sessions

Sabine Pfeiffer | bidt & FAU Erlangen-Nürnberg
Felix Freiling | bidt & FAU Erlangen-Nürnberg
Angela Graf | bidt
Andreas Wenninger | bidt